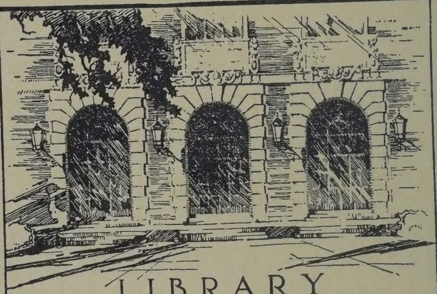


LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

DIGITAL
COMPUTER LAB.

Allerton Conference on Circuit
and System Theory 1965

Proceedings

7210
140

Proceedings

THIRD ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

October 20-22, 1965
University of Illinois
Allerton House, Monticello, Illinois

DIGITAL COMPUTER
LABORATORY
LIBRARY

Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois, Urbana, Illinois and the
Circuit Theory Group of the
Institute of Electrical & Electronics Engineers

OCT 25
1968

PROCEEDINGS
THIRD ANNUAL ALLERTON CONFERENCE
ON CIRCUIT AND SYSTEM THEORY

EDITED BY

M. E. VAN VALKENBURG
Conference Chairman

Conference held

October 20, 21, 22, 1965
Allerton House
Monticello, Illinois

Sponsored by the

DEPARTMENT OF ELECTRICAL ENGINEERING
and the
COORDINATED SCIENCE LABORATORY
of the
UNIVERSITY OF ILLINOIS
Urbana, Illinois

and the

CIRCUIT THEORY GROUP
of the
INSTITUTE OF ELECTRICAL & ELECTRONICS ENGINEERS

DIGITAL COMPUTER
LABORATORY
LIBRARY

FOREWORD

These Proceedings serve as a permanent record of the papers presented at the third in a series of conferences held at Allerton House devoted to the theoretical aspects of the study of circuits and systems. The Conference was founded on the premise that there is a unity in the field of circuit theory and system theory, that workers in the two fields have much to learn from each other, and that we need more forums for the presentation of ideas. As for past Conferences, Allerton House proved to be an ideal environment for discussions.

It is a pleasure to acknowledge the partial support of the sessions relating to sensitivity by the Air Force Office of Scientific Research under Grant AF-AFOSR 931-65, and the co-operation of Major B. R. Agins and Major B.S. Morgan, Jr. The Conference was co-sponsored by the University of Illinois and the Circuit Theory Group of the Institute of Electrical and Electronics Engineers.

We acknowledge the encouragement, support, and co-operation of Professor E. C. Jordan, Head of the Department of Electrical Engineering at the University of Illinois, and Professor W. Dale Compton, Director of the Coordinated Science Laboratory. Arrangements for the Conference were coordinated with extraordinary skill by Mrs. Divona Keel, and the business aspects of the Conference and these Proceedings were handled by Mr. Harold B. Lawler and Mrs. Rosa Townsend. Our colleagues participated in every phase of planning for the Conference; we are indebted to Professors Robert T. Chien, Wataru Mayeda, James A. Resh, Sundaram Seshu and Manoel Sobral, Jr.

J. B. Cruz, Jr.
W. R. Perkins
M. E. Van Valkenburg

TABLE OF CONTENTS

FOREWORD iii

GENERAL SESSION A 1
Chairman: M. E. Van Valkenburg, University of Illinois

Some Applications of Topology to Networks, Solomon Lefschetz 1

On Three-Terminal Circuits Without Transformers, Sidney Darlington 7

Scattering Techniques in Linear Systems, Herbert J. Carlin 14

Satisfaction Approach to the Synthesis and Control of Systems, Mihajlo D. Mesarovic 930

SENSITIVITY I 17
Chairman: William R. Perkins, University of Illinois

Controllability, Invariancy, Sensitivity, Rajko Tomović 17

On the Generalization of Sensitivity Equations, R. S. Rutman and P. V. Kokotović 27

Application of Game Theory to the Sensitivity Design of Systems with Optimal Controller Structures, P. Dorato and A. Kestenbaum 35

System Sensitivity in the Time-Domain, Richard C. Dorf. 46

SHIFT REGISTER CIRCUITS AND SWITCHING CIRCUITS 63
Chairman: Robert T. Chien, University of Illinois

Transfer Function Synthesis of Linear Shift-Register Circuits, D. T. Tang 63

A New Diagram for Feedback Shift-Registers, J. L. Massey and Ruey-Wen Liu 73

Cycle Structure of Monotone Feedback Shift Registers, Frederic J. Mowle and Daniel D. Sullivan 82

Sequential Codes and Discrete Memoryless Channels with an Erasure Property, Terry J. Wagner 91

Functional Transformations for Testing the Realizability of Single-Contact Functions, S. Park Chan and Edward R. Meyer. 99

STOCHASTIC OPTIMAL AND ADAPTIVE CONTROL 109
Chairman: J. Peschon, Stanford Research Institute

Combined Optimal Control and Estimation, Lewis Meier, III 109

Relative Costs of Mean and Variance Control for a Class of Linear, Noisy Systems, Michael K. Sain 121

Adaptive Control Using Time Delay and Correlation, Kumpati S. Narendra, M.A.L. Thathachar, and Theodore S. Baker 130

DISTRIBUTED NETWORKS 144
Chairman: B. A. Shenoi, University of Minnesota

Analysis of Distributed Parameter Networks - A General Method, Daniel Dicker 144

Frequency Characteristics of a Double-Kelvin Transmission Line Network, B. A. Shenoi 151

Distribution of Zeros of ABCD Parameters of Arbitrary RC Transmission Lines, E. N. Protonotarios and O. Wing 162

NETWORK SYNTHESIS STRATEGIES 172
Chairman: E. C. Jones, Jr., University of Illinois

The Generalization of the Miyata Realization Procedure, Hyong Kap Kim 172

The Application of Carroll's Optimization Technique to Network Synthesis, P. O. Scheibe and E. A. Huber . 182

Some Results Concerning Low-Pass Ladder Networks, F. F. Kuo 192

Ladder Network Synthesis by Similarity Transformation, Thomas G. Marshall, Jr. 200

A General Transformation Procedure in Circuit Theory, Marco A. Murray-Lasso 201

GENERAL SESSION B 202
Chairman: J. B. Cruz, Jr., University of Illinois

Sensitivity Analysis with Random Increments, A. Papoulis 202

SENSITIVITY II. 212
Chairman: Ronald A. Rohrer, State University of New York at Stony Brook

Some Approaches to the Reduction of Control System Sensitivity, P. Kokotović, S. Bingulac and J. Medanić 212

On the Minimization of Sensitivity in Optimal Control Systems, D. D. Siljak and R. C. Dorf 225

Minimization of Sensitivity with Stability Constraints in Linear Control Systems, D. D. Siljak and A. Burzio. 230

Some Theoretical Limitations of System Sensitivity Reduction, William A. Porter 241

Computational Procedures for Sensitivity Coefficients in Time-Invariant Multivariable Systems, Major Bernard S. Morgan, Jr. 252

STABILITY OF NONLINEAR SYSTEMS 259
Chairman: J. K. Aggarwal, University of Texas

Stability of Single-Loop Nonlinear Feedback Systems, Chung T. Lee and Charles A. Desoer 259

Stability Criteria of Phase-Locked Loops, P.W.U. Graefe and N. K. Loh 270

Analysis of Systems with Multiple Continuous Nonlinearities, Stuart A. Steele 282

The Exponential Describing Function in the Analysis of Nonlinear Systems, Theodore A. Bickart 283

SIGNAL THEORY 299
Chairman: S. Seshu, University of Illinois

On Bandlimited Properties of Fourier Transform Pairs of Some Special Functions, K. Yao and J. B. Thomas . . 299

Fourier Techniques for the Resolution of Overlapping Waveforms, J. R. Reeder and S. S. Wolff 310

Some Applications of Reproducing Kernel Hilbert Spaces, Thomas Kailath 320

Some Criteria for Frequency Domain Design of Signals Imbedded in Gaussian Noise, Marvin K. Simon and Ludwik Kurz . . . 329

OPTIMAL CONTROL . . . 339
Chairman: M. Sobral, Jr., University of Illinois

Optimal Open-Loop Control of Linear Systems with Time Lag, Thomas E. Mueller . . . 339

Controllability of Linear Systems with Time Lag, G. S. Tahim . . . 346

Controllability of a Class of Nonlinear Systems with Bounded State and Control Spaces, William D. O'Neill 353

Inequality Constraints in the Minimum Time Control Problem, N. B. Hemesath . . . 362

Optimal Design of Regulators, N. N. Puri . . . 369

SENSITIVITY III . . . 384
Chairman: M. E. Van Valkenburg, University of Illinois

Multiparameter Sensitivity Analysis for Linear Systems, R. N. Biswas and E. S. Kuh . . . 384

A Numerical Algorithm for the Minimization of Sensitivity, D. A. Calahan . . . 394

On Multiparameter Sensitivity, Samuel C. Lee . . . 407

On the Effect of Loss in Filter Networks, Maier L. Blostein . . . 421

Symmetrical Parallel-Tee Network--Parameter Plane Analysis and Synthesis, F. H. Hollister and G. J. Thaler . . . 430

MODERN SYSTEM THEORY . . . 439
Chairman: Gordon J. Murphy, Northwestern University

Decoupling of Multivariable Systems by Means of State Variable Feedback, Z. V. Rekasius . . . 439

Effective Construction of Linear State-Variable Models from Input/Output Data, B. L. Ho and R. E. Kalman . 449

Nonlinear Time Varying Feedback Systems--Conditions for L_{∞}-Boundedness Derived Using Conic Operators on Exponentially Weighted Spaces, G. Zames . . . 460

Recursive Estimation of the Modal Trajectory,
Robert E. Larson and John Peschon 472

Nonlinear Programming for Optimum Control, John Peschon . 482

Optimal Time-Domain Synthesis by Iterative Methods,
L. E. McBride, H. W. Schaefgen and K. Steiglitz . . . 492

N-PORT NETWORK THEORY. 502
Chairman: James A. Resh, University of Illinois

A Unified Theory of Linear N-Port Networks,
L. O. Chua and B. J. Leon 502

Restrictions on the Parameters of Resistive N-Port Networks, G. E. Sharpe 512

Synthesis of Grounded RC N-Ports by Canonical Star-Mesh Networks, J. E. Olivares, Jr. 513

Time Domain Synthesis of a Class of RLC Networks,
Donald M. Levy and David P. Brown 827

OSCILLATORS AND NON-LINEAR SYSTEMS 514
Chairman: G. H. Fett, University of Illinois

On Coupled Oscillators, J. K. Aggarwal and
C. G. Richie 514

Lower Bounds on the Thresholds of Subharmonic Oscillations, T. N. Trick and D. R. Anderson 524

On the Dynamical Equations of Nonlinear Networks with n-Coupled Elements, Carlos A. Holzmann and
Ruey-Wen Liu 536

A Method for Studying the Behavior of Nonlinear Systems Near and at Infinity of the Phase Plane,
Wayne H. Chen 546

GENERAL SESSION C. 563
Chairman: W.R. Perkins, University of Illinois

Algebraic Theory of Linear Systems, R. E. Kalman 563

SENSITIVITY IV 578
Chairman: J. B. Cruz, Jr., University of Illinois

Parameter Invariance Through the Use of Nonlinear Comparators, James M. Ham 578

On the Problem of Indeterminacy in System Identification, H. F. Meissinger 589

Sensitivity Comparison of Open-Loop and Closed-Loop Systems, J. B. Cruz, Jr. and W. R. Perkins 607

Sensitivity of Sampled Data Systems with Finite Sampling Width, L. Nardizzi 613

Some Practical Considerations in the Realization of Linear Digital Filters, J. F. Kaiser 621

NETWORK ANALYSIS AND DESIGN BY DIGITAL COMPUTERS 634
Chairman: F. F. Kuo, Bell Telephone Laboratories

Analysis and Design of Linear Networks with Adjustable Parameters Using On-Line Simulation, H. C. So . . . 634

An Iterative Method for Solution of Nonlinear Resistive Networks, Jacob Katzenelson and Leon H. Seitelman. 647

Comprehensive Active Network Analysis by Digital Computer - A State-Space Approach, Christopher Pottle 659

A Numerical Solution of Linear Time-Invariant Systems, M. L. Liou 669

Practical Filter Design Using Mathematical Optimization, A. D. Waren and L. S. Lasdon 677

Digital Computer Simulation of Communication Systems Using the Block Diagram Compiler: Blodib, Roger M. Golden. 690

FOUNDATIONS OF CIRCUIT THEORY 708
Chairman: F. T. Boesch, Bell Telephone Laboratories

A Distributional Study of the Real Frequency Behavior of Passive Systems, M. R. Wohlers 708

Immittances of Reduced Degree and Non-Positive Real Immittances in Network Synthesis, John A. C. Bingham 714

More About Good and Bad Two Element Kind Structures, Harry B. Lee 715

Outer and Inner Bounds on Driving-Point Impedance Functions of LLFP Networks, T. S. Huang and W. H. Ku 722

Non-Linear Circuit Modeling, Joseph R. Edenhofer 726

LINEAR GRAPH THEORY I 727
Chairman: D. T. Tang, IBM Research Laboratories

An Application of the 4-Color "Theorem" to Network Theory, Lee M. Maxwell and Richard E. Saeks. 727

Plane Representations and Codes for Linear Graphs, Louis Weinberg 733

An Asymptote for Large System Models, S. D. Bedrosian and J. Gulati. 745

Analysis of Lossy Communication Nets by Modified Incidence Matrices, Tadao Murata 751

Stochastic Flow and Efficient Transmission Through Communication Nets, Kenji Onaga. 762

Traffic Network Analysis, J. L. Goodnuff and W. B. Hall 772

FILTER DESIGN 773
Chairman: G. Szentirmai, Bell Telephone Laboratories

Iterative Chebyshev Approximation Technique for Equalizer Synthesis, G. C. Temes and J.A.C. Bingham 773

Some Considerations on Min-Max Approximation Problems for Network Design, Yasutoshi Ishizaki, Hiroshi Wakazuki and Hitoshi Watanabe. 786

Towards a Unified Theory for the Approximation of Specified Flat Bands by Network Functions, Mike Sablatash 796

Considerations for the Design of Crystal Filters, E. Christian and E. Eisenmann 806

An Approximation to a Specified Magnitude and Maximally Flat Delay Using Non-Minimum Phase Networks, S. Karni and Nasir Ahemd 817

Improved Active Realization of Elliptic-Function Filters, P. M. Lin 826

Ultraspherical Filters - Time and Frequency Responses, D. E. Johnson and J. R. Johnson 827

CONTROL SYSTEMS 828
Chairman: B.C. Kuo, University of Illinois

Suboptimal Control by Frequency Domain Techniques, Differential Approximation, and Quasilinearization, Andrew P. Sage 828

Determination of the State Variable A-Matrix by Direct Measurement, C. J. Triska and M. L. Bakker 840

Parameter Optimization in Linear Systems - A State-Space Approach, C. M. Bacon 841

Identification of System Parameters by Implicit Synthesis, S. E. Stephenson 842

The Cathode Ray Storage Tube in Adaptive Decision Systems, Baxter F. Womack and James H. Seamon . . . 843

CIRCUIT THEORY 844
Chairman: L. P. Huelsman, University of Arizona

Planar Equivalent Networks for Non-Planar Networks Derived by Matrix Methods, Erwin C. Hausmann, Gladwyn Lago, and Douglas H. Dobbs 844

An Equivalent Circuit for the Operational Amplifier, Edward M. Butler and Sanjit K. Mitra 845

A Classification of PR Biquadratic Immittance Functions and Their Realization Using a Perfectly Coupled and Center Branch Augmented Two-Port, Kurt H. Haase. . . 846

The Periodically Reverse-Switched Capacitor, John A. McKinney and Charles A. Halijak. 856

An Application of Residual Networks to Network Synthesis, L. B. Masten and R. H. Seacat 857

The Application of Signal Flow Graph Analysis to a Network Having a Time Varying Element, Rolly Hassun 867

LINEAR GRAPH THEORY II. 868
Chairman: W. Mayeda, University of Illinois

Topologically Distinct Subnets in a Complex Network, Hartley Jensen and J. W. La Patra 868

The Shortest Generalized Hamiltonian Tree, Narsingh Deo and S. L. Hakimi 879

Generation of Trees and k-Trees, Wai-Kai Chen 889

The Number of Trees in a Network, P. V. O'Neil and P. Slepian 900

The Number and Listing of All Trees in an Arbitrary Graph, B. R. Myers and L. V. Auth, Jr. 906

Algorithms for Determining Automorphism Groups for Planar Graphs, Louis Weinberg 913

SOME APPLICATIONS OF TOPOLOGY TO NETWORKS

SOLOMON LEFSCHETZ
Brown University, Center for Dynamical Systems, Providence, Rhode Island

ABSTRACT

Some recent reading on network theory suggests that in that theory one could utilize a few simple notions of one- and two dimensional topology. The aim of this talk is to present these notions and discuss various applications notably to Kirchoff's laws and to network duality.

Graph topology - Kirchoff's laws

You are certainly familar with graphs. A graph G consists of nodes (= vertices) $n_1, n_2, \ldots, n_{\alpha_0}$ and branches (= edges, arcs) $b_1, b_2, \ldots, b_{\alpha_1}$. Convenient assumptions are that G is connected and that every branch joins two distinct nodes.

The first important expression encountered is the Euler-Poincaré characteristic, or simply characteristic

$$x(G) = \alpha_0 - \alpha_1 . \tag{1}$$

Let x_k, y_k designate real numbers. With Henri Poincaré as prime mover topologists have been led to consider expressions called zero-chains and one-chains

$$C_0 = \Sigma\, x_k n_k \quad , \quad C_1 = \Sigma\, y_k b_k .$$

If b_k goes from n_j to n_k one defines the boundary of b_k as $\delta b_k = n_k - n_j$, written in the general form

$$\delta b_k = \Sigma\, \eta_{hs} n_s$$

$$\eta_{hs} = (0,\ldots,0,-1,0,\ldots,0,+1,0,\ldots,0),$$

and hence

$$\delta C_1 = \Sigma\, y_k \eta_k n_k . \tag{2}$$

In obvious vector-matrix notation (2) reads

$$\delta C_1 = y'b = y'\eta n . \tag{2'}$$

If $\delta C_1 = 0$, that is if C_1 has no boundary C_1 is a one-cycle or more simply a cycle. Cycles are characterized by

$$y'\eta = 0. \tag{3}$$

The $\alpha_1 \times \alpha_0$ matrix η is the incidence-matrix of the graph.

Simple example of a 1-chain: an oriented arc (fig. 1), and of a cycle: an oriented loop (fig. 2).

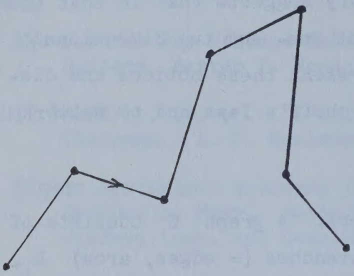


Fig 1.

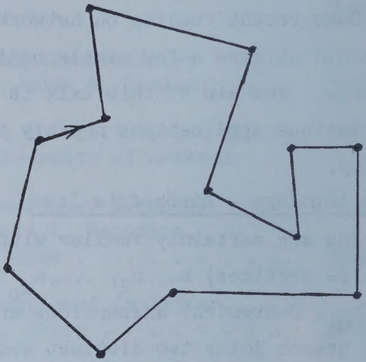


Fig. 2

The following is a short list of properties either readable directly from (3) or else readily proved otherwise.

I. Cycles make up a vector space of dimension $R = \alpha_1 - r$ where r is the rank of η. One calls R the Betti number of G (strictly "first Betti number of G).

II. $\chi(G) = 1 - r$.

III. There exists a base for the sysles made up of oriented loops $\lambda_1, \lambda_2, \ldots, \lambda_R$.

Theorem. Kirchoff's node (current) law means that every current distribution $i_1, \ldots, i_{\alpha_1}$ is a cycle: that is

$$i'\eta = 0 . \tag{4}$$

Let me give you a little more topology from a slightly different direction. For immediate convenience denote nodes and branches by $\overline{n}_1, \overline{n}_2, \ldots,$ and $\overline{b}_1, \overline{b}_2, \ldots$

Call zero-cochains and one-cochains the expressions

$$\overline{C}^0 = \Sigma x_k \overline{n}_k \ , \quad \overline{C}^1 = \Sigma y_k \overline{b}_k$$

and write associated coboundary relations

$$\partial \bar{n}_k = \Sigma \eta_{sh} \bar{b}_s$$

(expressing what branches end at n_k). Hence for a cochain coboundary

$$\partial \bar{C}^{0} = \Sigma x_h \eta_{sh} \bar{b}_s$$

or in vector-matrix form

$$\partial x' \bar{n} = x' \eta' b\ .$$

We are actually only interested in coboundaries and note:

IV. *Coboundaries form a linear vector space of dimension* α_1-R.

Theorem. Kirchoff's loop (voltage) law: a voltage distribution $v'\bar{b}$ *is a coboundary.*

That is $v'\bar{b} = \partial w'\bar{n} = w'\eta'\bar{b}$, and hence $v' = w'\eta'$, or $v = \eta w$

That is: there is a node voltage distribution $w'\bar{n}$ such that

$$\delta b_k = n_k - n_j \quad \text{implies} \quad v_k = w_k - w_j$$

(well known law).

A current i is a vector generating a subspace V_R of dimension R in a space V_{α_1} of dimension α_1. Similarily a voltage generates a subspace $V_{\alpha_{1-R}}$ of the same space. Now *these two subspaces are orthogonal*. That is

$$i'v = 0\ . \tag{5}$$

For $v = \eta w$, hence

$$i'v = i'\eta w = 0\ .$$

This very special case of a very well known theorem of topology is Telegen's theorem of network theory.

Observations.

The two results that we have given are well known to some (mostly mathematical) specialists. They provide the "idea trend" in regard to Kirchoff's laws. The vector-matrix formulations are readily turned into calculation procedures.

Note also that no linearity assumption has been required in the proofs.

Planar graphs.

The determination of the graph which may be sketched on a plane has been dealt with by three mathematicians: Kuratowski, Hassler Whitney

Saunders McLane. By utilizing slightly more advanced (but classical) algebraic topology I have obtained a rather simple proof of a theorem of Saunders McLane that I will now outline for you. It is necessary however to augument my dose of topology.

Let $\lambda_1, \lambda_2, \ldots \lambda_{\alpha_2}$ be a set of oriented loops of G. Cover λ_h with a 2-cell c_h (faithful image of a circular region), so that distinct e's are disjoint. The resulting figure K is a complex with α_o, α as before and with α_2 2-cells. The expression

$$\chi(K) = \alpha_o - \alpha_1 + \alpha_2$$

is the characteristic of K. The set of points in K is a polyhedron Π, also written $|K|$. Remarkably $\chi(K)$ depends solely an Π and on the special K covering it. Thus a cubical or tetrahedral surface cover some faithful reproduction of a sphere, and in fact their common χ is 2. We call it the characteristic of the sphere itself.

Noteworthy complexes are those covering a surface. The basic properties of such a complex are: (a) it is connected; (b) every branch is adjacent to exactly two cells; (c) the branches and cells ending at a given node n make up a sort of umbrella: branch, cell, branch,..., each branch adjacent to the cells that include it is this expression.

Theorem. The characteristic of a sphere is 2; that of any other surface is < 2.

Suppose that the graph G is planar and let it coincide with its image in a plane Π. Upon closing Π at infinity with a point it becomes a sphere S. Thus a planar graph may also be sketched on a sphere, and conversely.

Now as far as planting graphs in spheres one readily shows that one may restrict the search to a graph G with the following properties.

I. G is connected.

II. G is nonseparable: it has no node whose removal disjoins G.

IIIa. G is orientable: there are R + 1-$\mathcal{E}$ loops $\lambda_o, \lambda_1 + \ldots + \lambda_{R-\mathcal{E}}$ which satisfy just one relation

$$\lambda_o + \lambda_1 + \ldots + \lambda_{R-\mathcal{E}} = 0$$

and such that every branch belongs to exactly two loops of the set.

IIIb. We define G as nonorientable if it contains loops $\lambda_1, \ldots, \lambda_{R-\mathcal{E}}$ which satisfy no relation and such that every branch belongs to exactly two loops of the set.

Theorem of Saunders McLane. *A connected nonseparable graph with Betti number R is planar if and only if it possesses a set of loops $\lambda_1, \ldots, \lambda_R$ satisfying IIIa with $\varepsilon = 0$.*

Proof. If G is on a sphere the three conditions I,II,IIIa hold with $\varepsilon = 0$. Suppose now that I,II,IIIa with $\varepsilon = 0$ are satisfied. Construct K_G. Its characteristic is

$$\alpha_0 - \alpha_1 + \alpha_2 = 1 - R + (R + 1) = 2.$$

It is easily shown that if the umbrella condition fails $\chi < 2$, hence it must hold, K_G is a surface, consequently a sphere and the theorem follows.

When $\varepsilon \neq 0$ one must add the umbrella condition. Then according to the value of ε with or without orientability one has n.a.s.c. for immersion of a graph in a topologically uniquely defined surface.

Returning to planar graphs I only have time to mention that a standard dualization procedure of topology.(see fig. 3) is applicable. As far as I know it covers in full all dualization applied hitherto to networks.

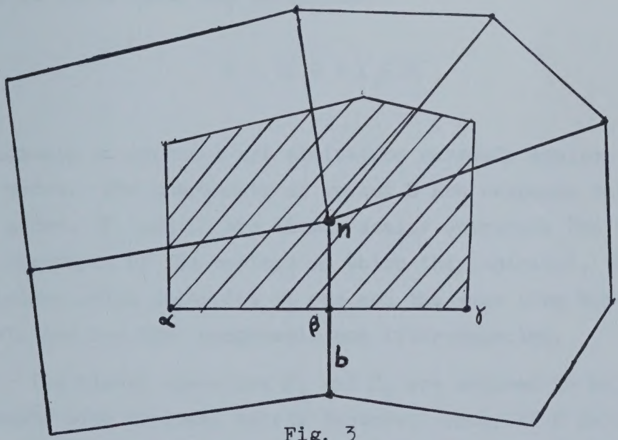


Fig. 3

Shaded cell dual to node n. Branch α β γ dual to branch b.

REFERENCES

1. K. Kuratowski, Sur le problème des courbes gauches en topologie, Fundamenta Matematicae (=F.M.), 15, 271-283, 1930.

2. Hassler Whitney, Nonseparable and planar graphs, Trans. Am. Math. Soc., 34, 331-362, 1932; F.M., 24, 23-34, 1933.

3. Saunders MacLane, A combinatorial condition for planar graphs, F.M., 28-32, 1936; A structural characterization of planar graphs, Duke Math. Journal, 3, 460-472, 1937.

4. S. Seshu and M. B. Reed, Linear graphs and electrical networks, Addison-Wesley, 1961. This volume has a very extensive set of references.

5. Further details on the proof of Saunders MacLane's theorem will be found in my paper: Planar graphs and related topics, Proceedings of the National Academy of Sciences, vol. 54, December 1965.

Supported in part by AFSOR under Contract No. 693-64.

ON THREE-TERMINAL CIRCUITS WITHOUT TRANSFORMERS

SIDNEY DARLINGTON
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

SUMMARY

The circuits discussed in this paper are interconnections of two kinds of linear, one-port components. Transformers and "coupled components" are excluded. Where it leads to a simpler discussion, the two component kinds are assumed to be positive, time-invariant resistors and capacitors. However, it has long been known that properties of "RC" circuits can easily be translated into properties of circuits using other pairs of component kinds. Furthermore much of the discussion is independent of the specific component kinds, provided the component "sizes" can be represented by constants required only to be non-negative, and provided the interconnections and sizes of the components of each kind can be represented by a matrix with the same properties as the node matrix of a circuit of resistors or capacitors.

In node matrix terms, a vector dynamical equation for circuit behavior in state space may be written

$$I = (\Gamma_1 G + \Gamma_2 C)E \tag{1}$$

The components of vector I are excitation currents applied to the various nodes. The components of vector E are response voltages at the various nodes. Γ_1 and Γ_2 are linear scaler operators (to be applied to each component of the vectors on which they operate). G and C are node matrices which determine at one and the same time the sizes of the components and how the components are interconnected.

The linear operators Γ_1 and Γ_2 are assumed to be such that they commute with constant matrix factors. Thus, if F is a vector and T is a constant matrix, Γ_1 and Γ_2 are to be such that

$$T\Gamma F = \Gamma T F \tag{2}$$

The matrices G and C are constrained to have the properties of node matrices of circuits of positive resistors without transformers. For a circuit of resistors and capacitors, Γ_1 is the identity operator, Γ_2 is the derivative operator, G is the node matrix for the resistors alone, and C is the node matrix for the capacitors alone.

The questions discussed relate to equivalent circuits (circuits which are internally different but which cannot be distinguished by external measurements). Most of the discussion concerns 3-terminal circuits (circuits in which only 3 terminals are accessible for external connections).

The principal analytical tool is the well known linear transformation on internal currents and voltages, as described for example by Howitt[1]. The linear transformation on the currents and voltages leads to a congruence transformation on each of the two node matrices, G and C, corresponding to the two component kinds. A real congruence transformation replaces a real matrix M by

$$\hat{M} = TMT^t \tag{3}$$

The transformation matrix T must be the same for both the component kinds.

Matrices G and C remain symmetrical and positive definite or semidefinite under transformations of the form (3). Very simple constraints on T keep the external properties of the 3-terminal circuits invariant. Interesting problems arise because the transformation must also keep all components non-negative. The most interesting transformations simplify the network topology by eliminating certain branches. But node matrices represent incomplete circuit graphs by assigning zero component sizes to the appropriate branches (edges) of complete graphs. Thus, interest centers on transformations which change certain component sizes to zero without making others negative. Furthermore the same transformation which eliminates some of, say, the resistors must yield no negative capacitors. Stated in other terms, one is interested in boundaries of the class of all transformations T which, when applied simultaneously to both the node matrices, yield no negative components.

The transformations T, by themselves, do not change the orders or ranks of the node matrices. On the other hand, many interesting equivalent circuits do have node matrices which are not of the same order or rank. However, useful changes in orders or ranks can be obtained by using some quite simple matrix manipulations in addition to the transformations T. Then one is interested in boundaries of the class of all transformations T plus means for changing orders and ranks which, when applied simultaneously to both node matrices, yield no negative components.

The purpose of this paper is two-fold: to disclose some properties of the circuits as such, and to promote the use of the transformations as a circuit theorist's tool. A little progress is reported toward solutions of still unsolved difficult problems concerning "classical" 3-terminal circuits. It illustrates the utility of general transformation principles which have been known and to some extent used by circuit theorists for over thirty years, but which have not enjoyed as extensive a use as they appear to deserve.

Mathematically, the techniques described are no more than applications of elementary matrix algebra. Each of the two node matrices (corresponding to the two component kinds) has the following properties. It is real and symmetrical. All elements on the main diagonal are positive or zero and all elements off the main diagonal are negative or zero. The sum of all the elements in any row or column is non-negative. It follows therefrom that the matrix is positive definite or semidefinite. Each diagonal element corresponds to a node and each nonzero off-diagonal element to a component. If $-e_{ij}$ is an off-diagonal matrix element, e_{ij} is the size of the corresponding component.

In the so-called indefinite form of the matrix every node in the circuit is represented by a diagonal element and every component is represented by an off-diagonal element. Conservation of currents requires that the sum of the elements in each row or column be zero and therefore the indefinite matrix is always singular. The so-called definite form requires the choice of some specific node as datum. It is derived from the indefinite form by eliminating the row and column representing components connected to the chosen datum. The sizes of components connected to the datum are now represented by sums over appropriate rows or columns in the reduced matrix.

It is convenient to partition a node matrix M as follows:

$$M = \begin{vmatrix} E & J^t \\ J & D \end{vmatrix} \qquad (4)$$

Here E corresponds to external nodes and components connected between pairs of external nodes. D corresponds to internal nodes and components connected between internal nodes or between internal nodes and datum. J and J^t correspond to components connected between external and internal nodes.

It is convenient to partition the transformation matrix T as follows:

$$T = \begin{vmatrix} U & B \\ 0 & A \end{vmatrix} \qquad (5)$$

Here U is the unit matrix of the same order as E in (4) (Order 2 when one of three external terminals is chosen as datum.), and 0 is a null matrix. Using (4) and (5) in (3) gives

$$TMT^t = \begin{vmatrix} (E + BJ + J^tB^t + BDB^t) & (J^tA^t + BDA^t) \\ (AJ + ADB^t) & ADA^t \end{vmatrix} \qquad (6)$$

One useful procedure is as follows: Choose A to modify D in some desired manner. Then choose B to obtain a more desirable transformed J. Finally, examine the properties of the transformed E. The only components determined by E are those connected between external terminals, and one can examine the transformed E to determine, for example, whether any of these "external" components have become negative. While this sort of circuit analysis is not entirely new, it appears to deserve more popularity than it has so far enjoyed.

The author published the following theorem some years ago, but without proof: Any n-terminal network of two kinds of positive components can be transformed into one in which each branch contains only a single component (never one of each kind), except for the branches terminating at pairs of external terminals.[2] The transformation may be accomplished as follows:

1. Choose an internal node as datum. Use a diagonal transformation matrix to adjust all branches between datum and internal nodes.
2. Repeat using each internal node as datum (in any order).
3. Repeat using each external node as datum.

At the same time, the author conjectured that the so-called "series parallel" configuration is canonical, within the class of all 3-terminal circuits of two kinds of positive components. It has by now been proved for circuits with only two internal nodes, but is in grave doubt for more complicated circuits. A proof for two internal nodes illustrates the following: Useful ways to increase the order of the node matrices, removal of a "small" T-circuit, and continuous transformation of a circuit as the limit of a sequence of small changes.

One way to increase orders and ranks is as follows: Replace matrix M by

$$\hat{M} = \begin{vmatrix} M & 0 \\ 0 & X \end{vmatrix}$$

where M is either G or C and X is diagonal and can be different for $\hat{G}$ and $\hat{C}$. When the components are resistors and capacitors, new matrix elements g_{ii} and c_{ii} add common factors $g_{ii} + sc_{ii}$ to numerators and denominators of the external admittance functions. An alternative procedure involves duplicating and rescaling an arbitrarily selected subnetwork within the circuit, in such a way that two subnetworks in parallel operate exactly like the original single subnetwork.

In a more recent publication, it was shown that every 3-terminal circuit of two kinds of positive components yields equivalent circuits of parallel T configurations, in which all components are positive except possibly components connected directly between external terminals.[3] In terms of the admittance functions associated with RC circuits, all components will be positive provided the frequency functions are sufficiently large near zero and infinite frequencies. The parallel T's are canonical unless there are other circuits with the same property except for smaller bounds on the behavior near zero and infinity. There frequently are such circuits and they raise the question of what configurations reduce to a minimum the price, in terms of behavior at zero and infinity, of realizing partial fractions of the frequency functions with given finite poles and residues.

At the same time, the author conjectured that parallel combinations of certain building blocks, including ladders as well as T's, are canonical. Under a rather simple condition the ladders are clearly excluded and then the conjectured theorem implies that the parallel T's can be used with no negative components. The condition requires all residues at poles of the transfer admittance function Y_{12} to be positive. The author has now proved the parallel T conjecture for two internal nodes, and for any number of nodes when Y_{12} is either zero at zero frequency or free of a pole at infinity. However he has disproved it for three internal nodes when Y_{12} is nonzero at zero and has a pole at infinity. The proofs illustrate also relations between the partial fraction analysis of the self- and driving-point admittance functions and transformation principles.

In connection with 2-terminal circuits of two kinds of positive components, specific transformations are known which transform any circuit either into a Foster canonical configuration or a Cauer canonical configuration. The transformation which yields a Foster configuration is easily adapted to 3-terminal circuits and leads to parallel T's (plus components between external terminals which may turn out to be negative). The transformation which yields a Cauer configuration can also be adapted to 3-terminal circuits. It leads to a fairly simple configuration, and all components are again positive except that negative components may again be required between external terminals. The conditions under which no negative components are required may be different for the two kinds of configurations.

Consider a circuit with more than one internal node. Suppose a single component (of one kind or the other) is connected from every internal node to every other node (internal or external), but that (initially) no components are connected between pairs of external nodes. Consider a transformation matrix T of the form

$$T = U + \varepsilon D$$

where U is the identity transformation and ε is an arbitrarily small scalar. There are matrices D with the following properties: The original configuration (one component per branch) is maintained, with small changes in component sizes (of order ε). The transformation may add small components between pairs of external terminals. They may be positive or negative, but are at most of order ε.

The above conditions do not determine D uniquely. The number of remaining disposable parameters is a function of the number of internal nodes, and possibly depends on other conditions in singular or degenerate situations. If there are enough disposable parameters, D can be so constrained that no components are introduced between external terminals. Then a sequence of small transformations can be followed until one or more of the original component values becomes zero.

Let n be the number of internal nodes and suppose $n \geq 4$. Then, except for possible troubles with singular matrices, a suitable sequence of transformations must reduce the 3-terminal network of two kinds of components to no more than 3n+6 "internal" components (all components not connected between pairs of external terminals). The more obvious effects of singular matrices leave the component count unchanged or else make it smaller. It has not been proved, however, that more subtle effects may not increase the minimum complexity.

REFERENCES

1. Howitt, N., "Equivalent Electrical Networks", PROC. IRE, Vol. 20 (June, 1932), pp. 1042-1051.

2. Darlington, S., "A Survey of Network Realization Techniques", IRE TRANS on CT, Vol. CT 2, (Dec., 1955), pp. 291-297.

3. Darlington, S., "Some Properties of Multiterminal RC Networks", IRE Int. Conv. Record, 1962, Part 2, pp. 36-48.

SCATTERING TECHNIQUES IN LINEAR SYSTEMS

HERBERT J. CARLIN, Electrophysics Department
Polytechnic Institute of Brooklyn
Farmingdale, New York 11735

SUMMARY

An important consideration in the application of scattering parameters to network problems is that these coefficients can be directly identified with physical quantities. Suppose a general electromagnetic structure is considered which consists of a cavity with metallic boundaries and n-ports for the entrance and exit of electromagnetic energy. The medium inside is presumed to be linear, but dispersive, inhomogeoenus, dissipative and anisotropic. It is then still possible to show that

1. If the medium is real frequency passive then

   $a^{+}(j\omega)(E - S^{+}(j\omega)\,S(j\omega))\,a(j\omega)$

   measures the real average power flow ($[\ ]^{+}$ means complex conjugate transpose, $S(j\omega)$ is the scattering matrix at real frequency ω, a is the vector of n incident wave amplitudes, and E is the unit matrix).
2. If the medium is lossless the scattering matrix $S(j\omega)$ is unitary.
3. These results are valid under real or complex normalization.

However, in a dissipative system it is only in special cases that the difference between average stored magnetic (W_H) and electric (W_E) energies can be related to the scattering parameters or for that matter any network representation. An example of a dissipative system in which ($W_H - W_E$) is meaningful is one with no dispersion which, in addition, is isotropic with losses that can be identified as purely ohmic. Under these conditions $a^{+}(S - S^{+})a = 2j\omega(W_H - W_E)$. This equation is therefore valid, for example, in a network containing R, L, C and transformers. The equation is not generally valid however in the non-reciprocal case e.g., if the network contains gyrators.

Because the scattering matrix relation to real power flow is valid under complex normalization one can show that the squared amplitude of an off diagonal scattering coefficient is equal to the

available gain between complex loads. Furthermore, the power relations show that the scattering matrix must exist for thermodynamically passive structures. Thus, networks which involve interconnected ideal transformers, or other frequency invariant systems such as multiport junctions, or circuits containing only resistors, gyrators, and transformers are quite tractable by S matrix techniques because the scattering representation exists for these circuits whereas the usual impedance representations generally fail.

In the case of a lossless medium which is defined in terms of dispersive tensor permeability and dispersive tensor dielectric constant it is again possible to obtain useful energy flow properties which are directly expressed in terms of the scattering parameters. In particular, it can be shown in a very general way that

$$j\, S^{+}(j\,\omega)\, \frac{d\, S(j\,\omega)}{d\,\omega}$$

is a positive hermitian matrix even under complex normalization to positive real functions. For normalization to real positive numbers this property has interesting physical consequences. For example, it can be shown that the average group delay for the two transmission directions, in a lossless 2 port resistively terminated, must always be non-negative, though in a non-reciprocal structure the delay function may actually be negative in one direction. Furthermore, the average delay function

$$-\frac{1}{2}\left(\frac{d\,\phi_{12}}{d\,\omega}+\frac{d\,\phi_{21}}{d\,\omega}\right)=\frac{\text{Average stored energy}}{\text{Incident Power}}$$

where $s_{ij} = |s_{ij}|\, \varepsilon^{j\,\phi_{ij}}$, may be interpreted as time of energy transport.

A consequence of this result is that in a lossless reciprocal 2 port this delay function must be non-negative in both directions and one may deduce a necessary condition that an insertion loss function

$$A(\omega) = \ell n \left| \frac{1}{s_{12}^{2}(j\,\omega)} \right|$$

be realizable as a minimum phase network. The condition is that

$$P \int^{\infty} \frac{x}{x^2 - \omega^2} \frac{d\,A(x)}{dx}\,dx \geq 0$$

where P denotes principle part.

REFERENCES

1. Carlin, H. J. and Giordano, A. B., "Network Theory: An Introduction to Reciprocal and Non-Reciprocal Circuits", Prentice Hall, 1964.
2. Carlin, H. J., "Energy Transport and the Scattering Matrix", Electronics Letters (British), May 1965.

Portions of the work described herein were sponsored by the Air Force Systems Command, Research and Technology Division, Rome Air Development Center under Contract No. AF 30(602)-2868 and by the Air Force Office of Scientific Research of the Office of Aerospace Research under Contract No. AF 49(638)-1373.

CONTROLLABILITY, INVARIANCY, SENSITIVITY

RAJKO TOMOVIC
University of Southern California
Electrical Engineering Department
Los Angeles, California

ABSTRACT

The controllability concept is studied in the light of perturbation properties of dynamical systems. This approach provides a general basis for the definition of controllability of both linear and nonlinear dynamical systems exposed to perturbations of piece-wise constant or time-varying parameters.

The perturbation properties of the dynamical system are expressed in the form of sensitivity functions. The amount of controllability of the dynamical system is determined by the minimization of a linear form of sensitivity functions corresponding to the selected set of parameters used for control purposes. The resulting controllability index lies in the range between 0 and 1 and measures the amount of system controllability, i. e. the distance between desired and actual states, for the given set of control parameters.

The perturbation properties of dynamical systems provide a common background for the understanding of mutual relations between the concepts of controllability, adaptivity and sensitivity.

1. INTRODUCTION

Controllability is concerned with the character of the set of reachable states of a given dynamical system. In controllability studies of linear systems the characteristics of the set of reachable states are considered irrespective of the concrete form of the control vector so that only intrinsic properties of the dynamical system are taken into consideration.

Originally the controllability concept was formulated for the set of all final states in the system state space. Such a study is fundamental for the solution of optimal control problems.

Recently the controllability concept was applied to the case when the set of reachable final states or controlled variables of the dynamical system has but one element1). This condition is essential for the design of so-called invariant systems2).

All the previous approaches, however, implied that the control vector which is instrumental in moving the dynamical system along a selected trajectory to the final state has the nature of a forcing function. In this contribution the control vector is treated in a more general way so that the motions of the dynamical system under variations of initial conditions, and piecewise constant and time-varying parameters, in addition to forcing functions, are studied in the light of the controllability concept. This generalization is essential in order

---

This work was performed while the author was a NSF Senior Foreign Scientist and Visiting Professor at the University of Southern California, Los Angeles.

to derive controllability criteria for adaptive systems.

In the past the controllability property has been treated primarily as an exclusive property of certain classes of dynamical systems. However a limited amount of controllability, i.e., the inherent property of dynamical systems to approach but not necessarily to reach the desired final state, is existent in all practical situations. Therefore, a controllability study explicitly taking the above fact into account would be very useful. This paper also undertakes such a generalization of the controllability concept.

2. THEORETICAL FOUNDATIONS OF CONTROLLABILITY STUDIES

Controllability criteria of linear dynamical systems can be conveniently derived by stating the problem in terms of matrix manipulations and the properties of linear vector spaces3). Another approach is to apply variational calculus. The latter approach was used to derive the uncontrollability criteria of dynamical systems1).

In order to generalize the scope of controllability studies, a theoretical approach closely associated with the sensitivity concept of dynamical systems is used here4). In other words the perturbation properties of dynamical systems under small variations of parameters are taken as a starting point in the development of the controllability concept and controllability tests. This approach leads to the necessary generalization of the control vector concept, gives a clearer insight into the physical background of controllability studies, and covers in a straight-forward way both non-linear and linear systems.

3. PERTURBATION APPROACH

Prior to analytical developments, the physical meaning of assumptions related to the present controllability study will be stated:

a) A controllability property is assumed to be shared to a continuously varying degree by all dynamical systems. This fact is expressed by associating a unique controllability index, defined over the continuous set of real numbers [0,1] to any given dynamical system.

b) Controllability is defined essentially as a local property of a dynamical system so that it covers both linear and nonlinear systems. The restricted class of systems and types of control vectors for which this local property can be extended globally will be stated explicitly.

The analytical and physical foundations of the perturbation approach to controllability rest in the so-called sensitivity functions. Since these functions form the basis for all further developments, a brief review of their main properties will be given here. If the reader is familiar with this approach, he can proceed to section 4.

Let the mathematical model of a dynamic system be

$$\dot{x} = g(x, p, t) \qquad (1)$$

where

g is an n-dimensional vector function
x(t) is the n-dimensional state vector

p(t) is an m-dimensional parameter vector including initial conditions.

The general case of deterministic perturbations of system (1) can be reduced to time-varying perturbations of time-varying parameters. It will be assumed therefore that the perturbed parameter vector $\overline{p}(t)$ is of the form $\overline{p}(t) = p(t) + \varepsilon \cdot v(t)$, where the components of v(t) are arbitrary, uniformly bounded, integrable functions and ε is a small constant parameter.

In order to give a rigorous meaning to the sensitivity function in the above case, the variational calculus will be used. Let us suppose that the functional is of the form

$$J[y] = \int_a^b F(y, \dot{y}, t)\, dt$$

Then the following equation is valid

$$\frac{\delta J}{\delta y} = F_y - \frac{d}{dt} F_{\dot{y}} \tag{2}$$

where $\frac{\delta J}{\delta y}$ is the functional (variational) derivative$^{5)}$. Its meaning is defined by the following expression

$$\left.\frac{\delta J}{\delta y}\right|_{t=t'} = \lim_{\Delta c \to 0} \frac{J[y+h] - J[y]}{\Delta c}, \quad t' \epsilon [a, b]$$

where h(t) is an increment of y(t) different from zero only in the neighborhood of t = t', and Δc is the area between the curve y = h(t) and t-axis. The area Δc goes to zero in such a way that both max $|h(t)|$ and the length of the interval in which h(t) is nonvanishing go to zero.

From (2) it is evident that the functional derivative represents the left-hand side of Euler's equation. In the case of an extremal value of the functional $\frac{\delta J}{\delta y} \equiv 0$, which is the analog of extremum conditions in differential calculus.

Equation (1) can be written in the form of an integral equation

$$x(p, t) = x(t_o) + \int_{t_o}^{t} g(x, p, \tau)\, d\tau . \tag{3}$$

The functional derivative $\frac{\partial x(p, t)}{\partial v(t)}$ can be defined with respect to the argument $\overline{p} = p(t) + \delta p(t)$, where δp is equal to $\epsilon \cdot v(t)$ and the matrix U is defined in the following way:

$$\frac{\partial x(p, t)}{\partial v} = \lim_{\epsilon \to 0} \frac{x(p + \epsilon v, t) - x(p, t)}{\epsilon} = U(t, p(t))$$

Using equation (2), the sensitivity equation of system (1) for time-varying perturbations is obtained

$$\dot{U} = G(\varphi, p, t) \cdot U + L(\varphi, p, t)$$

$$U(t_o) = 0 \tag{4}$$

Since x depends on p in the functional (3), the application of equation (2) requires the knowledge of the following matrices

$$F_x \triangleq \left\{ \frac{\partial g_i}{\partial x_j} \right\} = G(\varphi, p, t) \quad F_{\dot{x}} = 0 \quad i, j = 1, 2, \ldots n$$

$$F_{\bar{p}} \triangleq \left\{ \frac{\partial g_i}{\partial (p_k + \varepsilon \cdot v_k)} \cdot \frac{\partial (p_k + \varepsilon v_k)}{\partial \varepsilon} \right\} \quad F_{\bar{p}} = 0 \quad k = 1, 2, \ldots m$$

$$F_p = F_{\bar{p}} \Big|_{\varepsilon = 0} = \left\{ \frac{\partial g_i}{\partial p_k} \cdot v_k \right\} = L(\varphi, p, t)$$

Consequently

$$\delta J = F_x \cdot \delta x + F_p \cdot \delta p$$

which corresponds to equation (4).

Equation (4) reduces for constant parameters and constant perturbations to the usual sensitivity equation derived by differential calculus. Calculation of sensitivity functions requires the knowledge of system solution (1) so that in the general case they depend on system inputs and initial conditions. It should be kept in mind that sensitivity functions are defined only in the limit as ε → 0.

Sensitivity functions represent normalized local (differential) responses of dynamical systems to different kinds of perturbations. Reflecting as they do the intrinsic measurable properties of dynamical systems exposed to perturbations, sensitivity functions play an important role in control engineering in general. Moreover, their existence is not exclusively limited to analytical definition. In the so-called structural methods developed by Bihovski[4] and Kokotovic[6] it was demonstrated that sensitivity functions can be measured directly on dynamical systems without ever using mathematical models.

4. CONTROLLABILITY CRITERIA

We will state the definition of controllability before deriving controllability criteria. The usual definition of controllability will serve as a starting point[3]. It is reproduced here for convenience.

Definition 1: A dynamical system is said to be completely state controllable on $[t_o, t_f]$ if for given t_o and t_f any final state $x(t_f)$ can be attained starting with arbitrary initial conditions at $t = t_o$.

We shall make use of this definition keeping in mind that its validity will be tested in local regions S_φ around the unperturbed solution φ of System (1).

Using sensitivity functions, the following linear perturbation model can be defined.

$$\Delta x_{i_\ell} = \sum_{k=1}^{m} u_{ik} \cdot v_k \tag{6}$$

This perturbation approximates a Δx calculated from (1) only in the vicinity S_φ of each solution $\varphi' = (\varphi_1, \ldots, \varphi_n)$ of system (1). Evidently, $\lim \Delta x_i = \Delta x_{i_\ell}$ if $\lim \varepsilon = 0$.

In the subsequent text sensitivity matrices $U(t_f, t_o, \varphi, p(t))$ of system (1) will play an important role.

$$U = \begin{bmatrix} \dfrac{\partial\varphi_1}{\partial v_1} & \dfrac{\partial\varphi_1}{\partial v_2} & \cdots & \dfrac{\partial\varphi_1}{\partial v_m} \\ - & - & \cdots & - \\ \dfrac{\partial\varphi_n}{\partial v_1} & \dfrac{\partial\varphi_n}{\partial v_2} & \cdots & \dfrac{\partial\varphi_n}{\partial v_m} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ - & - & \cdots & - \\ u_{n1} & u_{n2} & \cdots & u_{nm} \end{bmatrix} \tag{7}$$

Excluding the singular values of parameters, the linear perturbation model (6) can be used to describe the behavior of the system (1) around any unperturbed value of the parameter vector. How far the linearization process can be extended is an important question, but it will not be dealt with here since it does not affect the results of this study. In other words the present controllability approach requires only the existence of a domain S_φ for which (6) is valid, but it is not necessary to know its limits unless practical computations are carried out.

Equation (6) can be used to derive the following set of algebraic equations

$$\begin{array}{l} u_{11}\Delta_1 + u_{12}\cdot\Delta_2 + \ldots + u_{1n}\Delta_m = b_1 \\ - \quad - \quad - \quad - \quad - \quad = \quad . \\ - \quad - \quad - \quad - \quad - \quad = \quad . \\ u_{n1}\Delta_1 + u_{n2}\Delta_2 + \ldots + u_{nm}\Delta_m = b_n\,. \end{array} \tag{8}$$

where the Δ_k ($= v_k$) represent constant perturbations of the respective parameters $p_k(t)$ and $b_1, \ldots, b_n$ are fixed numbers defining arbitrary points in the linearized state space.

If an arbitrary perturbation $\Delta x(t_f)$ in the vicinity S_φ of a reference (nonperturbed) trajectory $\varphi(t, p(t))$ is to be reached, the components $b_1, b_2, \ldots, b_n$ in (8) must be selected without restrictions. A necessary condition for this, assuming $n \leq m$, evidently is that the rows of (8) be linearly independent. Thus the following lemma holds true.

Lemma 1: Dynamical system (1) is completely state controllable in a local region S_φ around an unperturbed solution if the sensitivity functions of the state vectors with respect to selected parameter perturbations are linearly independent ($n \leq m$) throughout S_φ.

Since u_{ik} depends on the solution $\varphi(t, p(t))$ of system (1), the conditions of Lemma 1 must be verified in the case of nonlinear systems whenever a change in the system solution is involved. Note, however, that the sensitivity functions are obtained in every case without any assumptions regarding the linearity of the original system (1). The only approximation involved is that used in calculating the perturbed system response. This affects the necessary number of controllability tests but not the validity of Lemma 1.

5. CONTROLLABILITY INDEX

Equation (8) can be used to define a measure of controllability with respect to a selected set of system parameters. A controllability index, d, is defined as

$$d_j = 1 - \min_{\Delta_k} \sum_{i=1}^{n} \frac{1}{b_i^2} \sum_{k=1}^{m} (b_i - u_{ik}\Delta_k)^2 \qquad (9)$$

Having expression (6) in view, it is easy to see that the controllability index measures the Euclidean distance between the fixed (desired) points in the linearized space and the actual points in the same linearized space. In our case the controllability index d_j depends not only on the mathematical model (1) of the dynamical system but also on the choice of components of the parameter vector p. To each combination of system parameters there corresponds, in general, a different sensitivity matrix (7) and, accordingly, a different controllability index. In other words, each dynamical system has a finite set of controllability indices depending on the dimensionality of its parameter space.

Let us also point out that for a given choice of parameter perturbations the form (9) produces a unique controllability index. This does not imply, evidently, that the various controllability indices of the same dynamical system are identical. In fact this approach to controllability opens the possibility of comparing various control modes of the same dynamical system.

The following lemma is stated based on the above remarks.

Lemma 2: Every dynamical system has associated with it a controllability index vector $d = (d_1, d_2, \ldots, d_s)$ where s represents the number of different ways of choosing the control vector on the parameter space. Each component of d corresponds to a specific choice of the parameters used for control purposes. The value of all controllability indices $d_1, d_2, \ldots, d_s$ lies on the closed interval $[0, 1]$.

From the control engineering point of view the cases where the controllability index d_j in (9) takes on the extreme values 1 or 0 are of particular interest. The cases $n \leq m$, and $n > m$, corresponding to different relative dimensionalities of state and parameter spaces, will be treated separately in order to simplify the presentation which follows.

1. $n \leq m$

(a) If the rows of the sensitivity matrix (7) are linearly dependent, the minimal value of the quadratic form in (9) is 1 and, consequently, $d_j = 0$. The system is uncontrollable in the state space, i.e., the set of reachable states has but one element. However, the system is fully adaptive or invariant in parameter space which is a highly desirable feature in compensation problems.

If the rows of the sensitivity matrix are linearly independent then $d_j = 1$. The system is completely controllable in the state space but uncontrollable in the parameter space (nonadaptive).

(b) In the general case the minimization of the form (9) may produce any number on the interval $[0,1]$. In fact, this case has been studied extensively in mathematics under the topic of the generalized inverses for matrices$^{7)}$. The control engineering application of this concept was considered in connection with final value control systems$^{8)}$.

The basic idea is to decompose the vector space S_φ of the linear perturbation model (6) into two orthogonal subspaces S_a and S_b which correspond respectively to the null space and range space of the sensitivity matrix U defined by (7). Then the controllability index d_j of (9) represents that component of the Euclidean distance between the desired and actual state variables.

2. $n > m$

If the rows of the sensitivity matrix are linearly dependent, then $d_j = 0$, so that the system can be made fully adaptive. In all other situations d_j will belong to the open interval $(0,1)$.

Table I was prepared in order to give clearer insight into the controllability properties for the various cases. It shows the character of controllability indices depending on the dimensionalities of state and parameter spaces.

Table I

	$n = m$	$n < m$	$n > m$
d_j	$[1,0]$	$[0,1]$	$[0,1]$

It must be stressed that controllability tests using this approach are reduced to minimization of linear forms. Thus well developed standard computer algorithms for this purpose can be readily used.

6. LINEAR TIME-VARYING SYSTEMS

It can be shown that the perturbation technique provides the same results as the conventional approach in the case of linear systems$^{3)}$.

In this special case system (1) can be written as

$$x = A(t)\,x + b(t)\,u(t)$$

where

A(t) is a n χ n matrix
b(t) is a n vector
u(t) is the control variable (scalar)

The parameter vector used to control the system in this case is $p(t) = b(t)$ and the perturbations are equal to $b(t) + \varepsilon \cdot b(t)$.

The sensitivity equation (4) in this special case has the form

$$Z = A(t) \cdot Z + b(t) \cdot u(t) \cdot \delta_{ij} \tag{11}$$

where

$$z_j = \left[\frac{\partial x_1}{\partial b_j}, \frac{\partial x_2}{\partial b_j}, \ldots, \frac{\partial x_n}{\partial b_j} \right]', \quad j = 1, 2, \ldots, m, \ \delta_{ij} = 0, \text{ for } i \neq j, \ \delta_{ij} = 1, \text{ for } i = j$$

and all initial conditions of (11) are zero.

The solution of (11) is

$$z_j(t_f, t_o) = \int_{t_o}^{t_f} h_j(t_f, t) \cdot u(t) dt \tag{12}$$

where

$$h_j(t_f, t) = \Phi(t_f, t) \cdot b(t)$$

Note that $z_j(t_f, t_o)$ is the solution of the sensitivity equation when a perturbation is applied to the jth component of the parameter vector $p(t)$. By letting $u(t) = h_j'(t_f, t) \cdot \lambda$ where λ is fixed-length, nonzero vector the Gramian matrix

$$P_j(t_o, t_f) = \int_{t_o}^{t_f} h_j(t_f, t) \cdot h_j'(t_f t) \, dt$$

is obtained from which the conventional controllability criteria are easily derived[3]. Thus, as could be expected, local perturbation properties of linear dynamical systems with respect to vector $b(t)$ can be extended to global regions of state space.

The linear case is helpful for understanding the broad scope of controllability problems which can be treated by the perturbation approach. In particular the conventional approach to controllability now appears to be a rather restricted concept requiring both the mathematical model of the dynamical system to be linear and the perturbations of the parameter vector to be associated only with the coefficients $b(t)$ of the control variable $u(t)$.

7. CONCLUSION

Using perturbation techniques, the concept of controllability has been extended to the case when final states are attained by variation of the general parameter vector of the system. Thus, the close connection between controllability, invariancy and sensitivity has been established. This approach shows clearly that a necessary condition for nonlinear dynamical systems to be locally completely controllable is that they have nonzero, linearly independent, perturbation responses. In all other situations the dynamical system is controllable in a restricted sense. However, the amount of system controllability can be measured in a unique way allowing thus a convenient comparison either of various dynamical systems or of the same dynamical system under various modes of the control in its parameter space.

In addition this approach reduces controllability tests to minimization of linear forms involving sensitivity functions. Thus the general algorithms for the selection of extremal points in parameter space can invariably be used in all controllability tests.

ACKNOWLEDGMENT

The author is greatly indebted to Dr. N. E. Nahi, Professor at USC, for his valuable remarks and discussions during the preparation of this paper.

REFERENCES

1. Wang, P. K. C., "On Invariance, Uncontrollability and Unobservability in Dynamical Systems", USC Memorandum.

2. Preminger, J., Rootenberg, J., "Some Considerations Relating to Control Systems Employing the Invariance Principle", IEEE Trans., Vol. AC-9, No. 3, pp. 209-216, 1964.

3. Kreindler, E., Sarachik, P. E., "On the Concept of Controllability and Observability of Linear Systems", IEEE Trans. Vol. AC-9, No. 2, pp. 129-36, 1964.

4. Tomović, R., "Sensitivity Analysis of Dynamic Systems", McGraw-Hill, New York, 1964.

5. Gelfand, I. M., Fomin, S. V., "Calculus of Variations", Prentice-Hall, 1963.

6. Kokotovic, P., "Structural Method and the Sensitivity Theory of Automatic Control Ststems", Doctoral Thesis, Institute of Automation and Remote Control, Moscow, 1965.

7. Penrose, R., "A generalized inverse for Matrices", Proc. Cambridge Philos. Soc. 51 (1955).

8. J. R. Smith, N. E. Nahi, "Final Value Control Systems", IEEE Convention 1964.

ON THE GENERALIZATION OF SENSITIVITY EQUATIONS

R.S. RUTMAN
Institute for Automation and Remote Control ANSSSR
Moscow, USSR
P.V. KOKOTOVIĆ
Institute for Automation and Telecommunications
"Mihailo Pupin", Belgrade, Yugoslavia

ABSTRACT

A sensitivity operator is defined which represents the extention of the classical time-domain sensitivity function to a time-varying change of elements in a non-linear dynamic system. On the basis of implicit function theorems the generalized sensitivity equations are derived.

INTRODUCTION

The investigation of the effects of components variations on the behavior of control systems - the sensitivity problem - is recently receiving an increasing attention in the theory of automatic control. Although the concept of sensitivity has been so far insufficiently exploited in the classical feedback theory, it is as old as the feedback theory itself [1]. A thorough examination of the pertaining references[2] will disclose that the methods of error analysis of differential analysers and special procedures for the experimental derrivation and application of the sensitivity functions have given essential contributions to the sensitivity theory.

In accordance with the effort to extend the methods of automatic control on complex engineering, as well as biological and economic systems, attempts to generalize the concepts of sensitivity have recently appeared[3].Among others the "comparative sensitivity" for linear multivariable systems[4] and the definition of sensitivity through the weak derivative of a functional[5] have been introduced.

As a further step in the generalization of sensitivity, this paper defines the sensitivity operator as an extention of the concept of the sensitivity function to the case of small time-varying variations of elements in non-linear dynamic systems. To such a purpose, functional analysis may be conveniently utilized.

SYSTEM COORDINATES AND PARAMETERS

The behavoir of control systems is described through system coordinates and performance indices. Coordinates, as functions of time, define the state of the system at every instant of time, while the indices as functionals of system coordinates supply the information about the quality of the system in a more condensed form.

Coordinates and performance indices are elements x of a functional space X, $x \in X$. Let R be the space of all input quantities r, and Q the space of system parameters q and form the set $Q \times R$ of all possible pairs (q,r). The control system will be specified through a non-linear operation S mapping the set $\Omega \in Q \times R$ into the space X

$$x = S\ ((q,r)) \tag{1}$$

Operation S may be considered as an abstract function of two variables to which the methods of elementary theory of functions may easily be applied[6], and we shall write $S\ (q,r)$ instead of $S\ ((q,r))$. It should be pointed out that the operation S is usually given only implicitly through some other operation P, i.e. in the general case the control system is defined through the equation

$$P\ (q,r,x) = 0 \tag{2}$$

If the transformation P is continuous in the vicinity of the point $(q_o, r_o, x_o) \in Q \times R \times X$ and if all the other conditions for the existence of implicit functions are satisfied, the operation S is uniquely defined by the following relations:

a) $P(q,r,\ S(q,r)) = 0$

b) $S(q_o,r_o) = x_o$ (3)

c) S is continuous at the point x_o

The functional equation (2) may represent a set of differential equations and operation S the expression for the analytical dependence of the solution x on the parameters q and inputs r.

SENSITIVITY OPERATOR

In some special cases such as with quadratic integral performance indices for linear systems with constant parameters, it is possible to express the operation S explicitly. However, in any more complicated systems, the investigation of the dependence of coordinates and indices on parameters through the expression (1) is practically impossible. This brings forth the necessity to approximate (1) in the vicinity of the point $(q_o,\ r_o)$ for which the solution x_o of expression (2) is known.

The point $(q_o,\ r_o)$ is the "nominal point" and the solution x_o is the "nominal motion" of the system. The problem we shall now discuss is the approximation of the "additional motion" Δx due to a small variation Δq of the parameter q in the vicinity of the nominal point $(q_o,\ r_o)$ and for a given $r = r_o$. Omitting in the further text the dependence on the input quantities r we write $S(q_1) = x_1$, where $x_1 = x_o + \Delta x$ is the "perturbed motion" and $q_1 = q_o + \Delta q$.

If the terms of higher order in Δq are neglected in the expansion

$$\Delta x = S(q_o + \Delta q) - S(q_o) = U(\Delta q) + O(\|\Delta q\|)$$

the additional motion Δx may be expressed through a quantity $U(\Delta q)$ obtained by applying the operation U to the variation Δq :

$$\Delta x \approx U(\Delta q) \qquad (4)$$

The linear operation U represents the Frechet derivative of the non-linear operation S at the point q_o, $U = S'(q_o)$, and characterizes the dependence of the system behaviour on the parameters[6,7]. Analogous to the sensitivity functions U will be called the "sensitivity operator".

The sensitivity operator U may be obtained directly from the operation P without the explicit knowledge of the operation S. This possibility represents the foundation upon which all the procedures of the sensitivity theory are based and is expressed through the theorem of the derivative of abstract implicit functions:

$$U = -\left[P'_x /q_o, x_o/\right]^{-1} P'_q /q_o, x_o / \qquad (5)$$

where

$P'_x(q_o, x_o)$ and $P'_q(q_o, x_o)$ are partial derivatives with respect to x and q of the operator P at the point (q_o, x_o). The additional motion Δx may be obtained approximatively if the operations P'_q and then the inverse of the linear operation P'_x are applied to the parameter variation Δq:

$$\Delta x = U(\Delta q) = -\left[P'_x/q_o,x_o/\right]^{-1} P'_q/q_o,x_o/(\Delta q) \qquad (6)$$

The algorithm for obtaining the additional motion is illustrated in the block diagram Fig. 1.

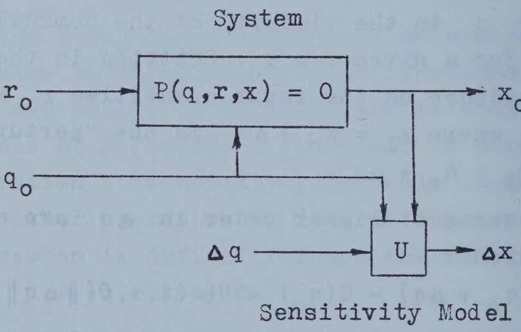


Fig. 1

In the formation of the operator U, quantities x_o and q_o may be considered as parameters and Δq as the input variable.

The relation (6) may be rewritten in the form of a generalized "sensitivity equation":

$$P'_x /q_o, x_o/(\Delta x) = - P'_q /q_o, x_o/(\Delta q) \qquad (7)$$

If Δq_j is a constant scalar quantity and Δx an n-dimensional state vector, by dividing (7) with Δq_j and passing to the limit, we obtain the classical sensitivity equations[6,7]

$$P'_x(q_o, x_o) \left(\frac{\partial x_o}{\partial q_j}\right) = - P'_q(q_o, x_o) \qquad (8)$$

where $\frac{\partial x}{\partial q_j}$ is one-parameter sensitivity vector.
If q is an m-dimensional constant vector, the procedure may be repeated for each of its components, and a system of equations defining the nxm sensitivity matrix

$$\left[\frac{\partial x_i}{\partial q_j}\right]; i = 1, \dots n; \quad j = 1, \dots m$$

may be obtained. Hence, for a fixed variation of Δq, the additional motion will be defined through the matrix equation

$$\Delta x = \left[\frac{\partial x_i}{\partial q_j}\right] \left[\Delta q\right]. \qquad (9)$$

Comparing (9) with (4) we see that for this case the operator U is represented by a sensitivity function matrix.

EXAMPLE

Find the sensitivity operator for the system represented by the equation

$$P(q, x, r) = \left[D + q\right] x + r = 0 \qquad (10)$$

D is the operator defining time differentiation, and q,r are functions of time. Taking into account the definition of the strong derivative of a non-linear operator, we have:

$$P'_x /q_o, x_o/ = D + q_o \quad , \qquad (11)$$

$$P'_q /q_o, x_o/ = x_o$$

and we obtain from (5)

$$U = - \left[D + q_o\right]^{-1} x_o . \tag{12}$$

The additional motion is defined by

$$\Delta x = - \left[D + q_o\right]^{-1} x_o \Delta q , \tag{13}$$

or in the form of a sensitivity equation

$$\left[D + q_o\right] \Delta x = - x_o \Delta q . \tag{14}$$

The equation may be represented in the form of a block diagram, Fig. 2.

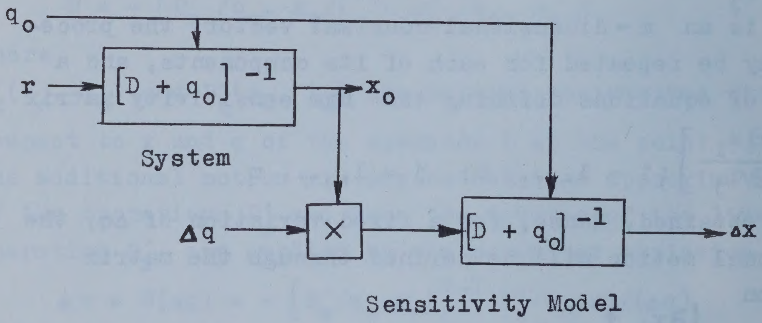


Fig. 2

It should be mentioned that in this simple example the operation S may be explicitly derived and the operator U obtained by simple derivation of the operation S.

CONCLUDING REMARKS

In the case of constant variations of parameters, the additional motion may be obtained by multiplying the sensitivity functions by the corresponding parameter variations. However, in time-varying systems the additional motion will depend on the form of the small, but time-varying parameter changes, and to obtain the additional motion the sensitivity operator must be applied. To answer the question if all the benefits of the sensitivity method are lost in this general case, the operator U and operation S should be compared. Evidently, the linear operator U may be analysed more easily than the non-linear operation S implicitly defined through the operation P.

Expression (6) points out the justification for the separate investigation of the additional motion, since in obtaining Δx on a computer or in some experimental way, the parameter variation may be applied in an enlarged scale due to the linearity of operator U. The additional motion may then be obtained with the same degree of accuracy as the nominal motion itself.

As the separation of parameters and input signals is arbitrary in complex systems, the above considerations may in full be extended on the variation of input magnitudes.

The authors wish to express their gratitude to Mihael Epelman for his advice in formulating this paper.

REFERENCES

1.- Bode H.W. NETWORK ANALYSIS AND FEEDBACK AMPLIFIER DESIGN. Van Nostrand, Princeton,N.Y., 1945.

2.- Kokotović P., Rutman R. SENSITIVITY OF AUTOMATIC CONTROL SYSTEMS. Automation and Remote Control, Vol. XXVI, No. 4, Moscow, April 1965, pp.730 - 746.

3.- Proceedings of the I ETAN/IFAC Symposium on Sensitivity Analysis, Dubrovnik 1964, Pergamon Press 1965.

4.- ibid.
Perkins W.R., Cruz J.B., Jr. SENSITIVITY OPERATORS FOR LINEAR TIME-VARYING SYSTEMS

5.- ibid.
Rissanen J. DESIGN OF SELF-ADJUSTING CONTROL SYSTEMS BY USE OF A FUNCTIONAL DERIVATIVE TECHNIQUE

6.- Kantorović L., Akilov G. FUNCTIONAL ANALYSIS IN NORMED SPACE (in Russian). FM, Moscow, 1959.

7.- Sarachik Ph.E. FUNCTIONAL ANALYSIS IN AUTOMATIC CONTROL, IEEE International Convention Record, New York, 1965.

APPLICATION OF GAME THEORY TO THE SENSITIVITY DESIGN OF SYSTEMS WITH OPTIMAL CONTROLLER STRUCTURES

P. DORATO and A. KESTENBAUM
Department of Electrical Engineering
Polytechnic Institute of Brooklyn
Graduate Center
Farmingdale, New York

ABSTRACT

The theory of games is applied to of design of systems with unknown plant parameters. It is assumed that a controller structure is known and furthermore this controller is optimal when the controller parameters are equal to the plant parameters. The performance index then becomes a function of plant and controller parameters. This function is treated as a pay-off function with the antagonists represented by the controller and plant parameters.

1. INTRODUCTION

The sensitivity problem in engineering arises from a basic lack of knowledge concerning the value of certain required parameters. A classic approach to the sensitivity problem has been to assume nominal values for the unknown parameters and then to study the effects of small variations about these nominal values via sensitivity coefficients or sensitivity functions. With a few exceptions[7, 8], the sensitivity problem in optimal control has been largely neglected. Recently game-theoretic notions have been applied to the sensitivity problem in optimal control by Drenick and Dorato[3] and Rohrer and Sobral[4]. For this class of problems, i. e., optimal-control sensitivity problems, the game-theoretic approach appears to be a very natural one and has the added advantage that parameter variations need not be limited to small variations. In the first paper cited above the controller structure is not pre-specified, that is an open structure control problem is assumed. For continuous-time problems this leads to the rather complex problem of games defined over a function space. In the present paper only fixed controller-structure problems are considered, in particular, the type of fixed-structure problem considered coincides with the problem developed in the example of the second paper cited above. Specifically, the following problem is considered here:

Given a closed-loop control law

$$\underline{u}(t) = \underline{\psi}(\underline{x}(t),\ w_c;\ t) \tag{1}$$

which is optimal, when $w_c = w_p$, for a plant with dynamics

$$\dot{\underline{x}}(t) = f(\underline{x}(t),\ \underline{u}(t),\ w_p) \tag{2}$$

and a performance index (which is minimized)

$$S = \int_{t_o}^{T} F(\underline{x},\ \underline{u})\, dt \tag{3}$$

determine a "best" value of w_c when all that is known about w_p is that it lies somewhere in the range $w_1 \leq w_p \leq w_2$. For the sake of simplicity the parameters w_p and w_c are assumed to be scalars, all other variables, i. e., $\underline{x}(t)$, the state of the plant and $\underline{u}(t)$, the control input, may be taken to be vectors. The above problem is the basic sensitivity design problem to be considered in this paper. The sense in which a particular value of w_c is best, is developed in the next section.

2. SOLUTION OF THE SENSITIVITY DESIGN PROBLEM

With the assumptions of the previous section the performance index S becomes a function of the two variables w_c and w_p, i. e.,

$$S = S(w_c, w_p)$$

and for a given w_p the value of w_c which minimizes S is again by assumption, $w_c = w_p$. The problem is then to choose a value of w_c which maintains as low as possible a value of S given that w_p may lie anywhere in the range $w_1 \leq w_p \leq w_2$. It is assumed here that w_c is also constrained to the closed interval $[w_1, w_2]$. From the assumption on the optimal structure of the controller, it seems reasonable to constrain the controller parameter to range over the same set of values as the plant parameter. This latter assumption however is not necessary for the development that follows, although it does simplify some of the computations that are ultimately required.

In light of the uncertainty in the value of w_p it would then seem desirable to choose a value of w_c^o, which maintains S below some value, denoted S^o, i. e.

$$S(w_c^o, w_p) \leq S^o, \text{ for } w_1 \leq w_p \leq w_2 \tag{4}$$

On the other hand if S^o is to have any meaning as far as keeping the value of S as low as possible is concerned, it should follow that for any value of w_c other than w_c^o there may exist values of w_p which causes S to exceed S^o. This latter statement is summarized by the inequality

$$S^o \leq S(w_c, w_p^o), \text{ for } w_1 \leq w_c \leq w_2 \tag{5}$$

Inequalities (4) and (5) have a direct game theoretic interpretation. Indeed S is referred to as the value of the game as represented by the "pay-off function" $S(w_c, w_p)$ and the "antagonists" w_c (which attempts to maximize S). The pair (w_c^o, w_p^o) is referred to as the "optimal (pure) strategy". For the sensitivity design problem only the value of w_c^o is of interest. Thus the sensitivity design for the class of problems considered here is reduced to a problem in the theory of games, in particular since S is defined over real and continuous arguments the type of game involved is the so-called "continuous" or "infinite" game.

One might object to the game-theoretic approach as being too pessimistic, that is too overly concerned with the worst that can happen. This is certainly true and represents a limitation of the game-theoretic approach. Quite often with a little more effort one can either gather more information* about the variation in w_p or one can introduce additional control loops to stabilize the value of w_p about some nominal setting. However there is some value in knowing what can be done without additional information or control, and this type of information is indeed supplied by the game-theoretic approach.

In the next section various pertinent definitions and results from the theory of games are presented. In the final section the game-theoretic approach suggested here is applied to several optimal control problems.

3. THEORY OF CONTINUOUS GAMES

The definitions and results presented in this section are standard and are presented only for the reader's convenience. Proofs may be found in the references cited and will be omitted here.

Definitions[1, 2]

(1) Antagonists The players in a game. Generally, one associates variables with each player, e. g., w_c, w_p, etc.

(2) Strategy A specification of how each player should play. In terms of the variables w_c and w_p, strategies are specified in terms of probability distributions, e. g., $F_c(w_c)$ and $F_p(w_p)$. If a distribution $F(w)$ is a unit step function, i. e, if a particular value of w is played with probability one, then the strategy is said to be pure, otherwise, it is said to be mixed. A particular case of a mixed strategy where $F(w)$ is a step-like function is referred to as a strategy of the finite type. If every value of w is played with some positive probability the strategy is said to be completely mixed.

(3) Pay-off Function A function of the player variables w_c and w_p, which represents the prize or gain associated with the game. Let this function be denoted $S(w_c, w_p)$.

(4) Game A game is basically represented by the payoff function and player strategies. If the set of possible values of the variables w_c and w_p are real, discrete, and finite, then the game is said to be finite (also discrete, rectangular) and the pay-off function may be represented as a matrix. If the set of possible values of w_c

---

* One might, for example, attempt to determine a probability func- for the distribution of the values w_p.

w_p represent a continuum on the real line, then the game is said to be infinite (also continuous). A game is said to have a solution if there exists distributions $F_c^o(w_c)$ and $F_p^o(w_p)$ such that

$$E_{F_c^o F_p}\left\{S(w_c, w_p)\right\} \leq E_{F_c^o F_p^o}\left\{S(w_c, w_p)\right\} \leq E_{F_c F_p^o}\left\{S(w_c, w_p)\right\}$$

where typically $E_{F_c F_p}\left\{S\right\}$ denotes the expected value of S with respect to the distributions F_c and F_p, i. e.,

$$E_{F_c F_p}\left\{S(w_c, w_p)\right\} = \iint S(w_c, w_p) dF_c(w_c) dF_p(w_p)$$

the strategies $F_c^o(w_c)$ and $F_p^o(w_p)$ are referred to as the optimal strategies and the expectation evaluated for the optimal strategies, i. e., the middle term in the above inequalities, is referred to as the value of the game and is often denoted val S. If the optimal strategies are pure then the above inequalities simplify to

$$S(w_c^o, w_p) \leq S(w_c^o, w_p^o \leq S(w_c, w_p^o)$$

(5) Minmax and Maxmin Strategies The strategies F_c^M and F_p^M are said to be minmax strategies if

$$E_{F_c^M F_p^M}\left\{S(w_c, w_p)\right\} = \min_{F_c} \max_{F_p} E_{F_c F_p}\left\{S(w_c, w_p)\right\}$$

The strategies F_c^m and F_p^m are said to be maxmin strategies if

$$E_{F_c^m F_p^m}\left\{S(w_c, w_p)\right\} = \max_{F_p} \min_{F_c} E_{F_c F_p}\left\{S(w_c, w_p)\right\}$$

(6) Constant Risk (Equalizer) Strategy A strategy $\tilde{F}_c(w_c)$ is said to be a constant risk or equalizer strategy if

$$E_{\tilde{F}_c}\left\{S(w_c, w_p)\right\}$$

is independent of w_p.

Results

Below are summarized some of the more useful results available in the theory of continuous games. Proofs of the results may be found in the literature cited.

(1) Fundamental Theorem for Continuous Games. (ref. 1, p. 186) If $S(w_c, w_p)$ is a continuous function of two variables in a closed unit square[*], then both $E_{F_c^M F_p^M}\{S\}$ and $E_{F_c^m F_p^m}\{S\}$ exist and are equal. Thus with the above restrictions optimal strategies, possibly mixed however, always exist.

(2) Minmax Inequalities (ref. 2, p. 8) For every game for which minmax $(w_c^M w_p^M)$ and maxmin (w_c^m, w_p^m) pure strategies exist it follows that

$$S(w_c^m, w_p^m) \leq S(w_c^M, w_p^M)$$

Also

$$S(w_c^M, w_p) \leq S(w_c^M, w_p^M), \text{ for } w_1 \leq w_p \leq w_2$$

so that when w_c^M is played a pay-off level $S(w_c^M, w_p^M)$ is always guaranteed. Finally, the value of the game always lies between $S(w_c^m, w_p^m)$ and $S(w_c^M, w_p^M)$.

(3) Conditions for Pure Optimal Strategies (ref. 2, p. 8) Below are two necessary and sufficient conditions for optimal strategies to be pure.

(i)

$$\min_{w_c} \max_{w_p} S(w_c, w_p) = \max_{w_p} \min_{w_c} S(w_c, w_p).$$

The value of the game is given by the common value above.

(ii) there exists numbers w_c^o, w_p^o, and S^o such that

$$S(w_c^o, w_p) \leq S^o \leq S(w_c, w_p^o)$$

The value of the game is given by S^o. From the assumption that the controller has an optimal structure, it follows that for the games which arise in this study, if w_p^o is an optimal strategy then $w_c = w_p^o$ must also be an optimal.

(4) Equalizer Strategies (ref. 2, p. 28) If F_p^o is a completely mixed strategy, then every optimal strategy F_c^o, pure or mixed, is an equalizer strategy[**].

---

[*] Note that the set $[w_1, w_2]$ can always be mapped into set $[0, 1]$. Hence, one might just as well talk of the unit square.

[**] The interesting thing about equalizer strategies as far as the sensitivity problem is concerned is that they manifest complete insensitivity.

(5) Games with Convex Pay-off Functions (ref. 1, chapter 12) A pay-off function is said to be convex with respect to w_c if

$$\frac{\partial^2 S}{\partial w_c^2} \geq 0, \text{ for } w_1 \leq w_p \leq w_2$$

and concave with respect to w_p if

$$\frac{\partial^2 S}{\partial w_p^2} \leq 0, \text{ for } w_1 \leq w_c \leq w_2$$

If S is convex in w_c then w_c possesses a pure strategy. If S is concave in w_p then w_c possesses a finite strategy with at most two steps.

(6) Games with Separable Pay-off Functions (ref. 1, chapter 11) A pay-off function is said to be separable if it can be reduced to the form

$$S(w_c, w_p) = \sum_{i=1}^{n} r_i(w_c)\, t_i(w_p)$$

If S is separable then every mixed strategy is equivalent to a finite strategy with, at most, n steps.

It should be noted that it is certainly more desirable, from the implementation point of view to have pure strategies for the controller parameter. Unfortunately this is not always possible, indeed mixed strategies may not even turn out to be of the finite type. The implication of mixed strategies for a controller parameter is that when the controller is realized parameter values must be chosen from some random experiment, for example, from a flip of a coin. In addition to the implementation difficulty, it is extremely difficult to obtain solutions in the mixed strategy case. Thus one must often content oneself with something less than a guaranteed performance index level represented by the val S. Result (2) indicates that if one is content with the level represented by $S(w_c^M, w_{p_M}^M)$, i.e., a minmax solution, then the pure strategy represented by w_c^M always suffices.

In the examples of the next section it develops that all solutions involve pure strategies only.

4. EXAMPLES

Example 1

Consider a plant with dynamics

$$\ddot{x} + w_p \dot{x} = u$$

with $x(0) = 1$, $\dot{x}(0) = 0$, and $0 \leq w_p \leq 2$ and a performance index

$$S = \int_0^\infty (x^2 + \dot{x}^2 + u^2)\,dt$$

The optimal controller structure for this problem is[(4)]

$$u^o = -x + (w_c - \sqrt{3 + w_c^2})\,\dot{x}$$

and the value of S is[(4)]

$$S(w_c, w_p) = \frac{(w_c - \sqrt{3 + w_c^2}) + 3}{2(w_p - w_c + \sqrt{3 + w_c^2})} + w_p$$

Let $c_2 = w_c - \sqrt{3 + w_c^2}$ and consider S in terms of the variable c_2 and w_p, i. e., let $S = S(c_2, w_p)$, where c_2 ranges from -1. 73 to -. 64. Figure 5 of reference (4) suggests that a minmax solution is given by the pair $c_2 = -.64$ (which corresponds to $w_c = 2$) and $w_p = 2$. (Note that in figure 5 of reference (4) the arguments of S are interchanged). If the optimal strategies are pure, the minmax solution is the optimal solution and the minmax value of S is the value of the game, i. e., val $S = S(-.64, 2) = 2.65$. One need then only verify the inequalities

$$S(-.64, w_p) \leq 2.65 \leq S(c_2, 2)$$

From figure one it is clear that these inequalities are indeed satisfied, hence the optimal controller parameter is $w_c = 2$. In terms of $S(w_c, w_p)$ this value of w_c guarantees that

$$S(w, w_p) \leq 2.65, \text{ for } 0 \leq w_p \leq 2$$

This same problem is worked out in reference (4) where the optimal value of c_2, based on a relative-sensitivity type of pay-off function, was found to be $c_2 = -1.3$. For this value of c_2 the best one can guarantee for S is about 2. 72. In addition the total units of variation for the relative sensitivity* approach is larger than the

---

* Relative sensitivity S^R as defined in reference (4) is given by

$$S^R = \frac{S(w_c, w_p) - S(w_p, w_p)}{S(w_p, w_p)}$$

Thus with S^R as a performance index the design object is to maintain $S(w_c, w_p)$ as close as possible to the "optimal" S-curve, $S(w_p, w_p)$.

variation given here, .95 units compared to .67 units. Of course, the choice $c_2 = -1.3$ does maintain the $S(-1.3, w_p)$ - curve close to the optimal S-curve. In the final analysis the criterion chosen will depend on the particular factors the designer considers most important.

Example 2

Consider a plant with dynamics

$$\ddot{x} = w_p u \,, \quad |u| \le 1$$

with $x(0) = 1$, $x(0) = 0$, and $.8 \le w_p \le 1.2$. The performance index is this problem is the time T taken to drive the system to the origin, i.e., $x(T) = 0$, $\dot{x}(T) = 0$.

The optimal controller structure in this case is known to be[6]

$$u = -\operatorname{sgn}\left(\frac{\dot{x}\,|\dot{x}\,|}{2w_c} + x\right)$$

If $w_p = w_c$, the system is optimal and the origin is reached after one switching. The value of T in this case is easily shown to be given by[6]

$$T = S(w_p, w_p) = \frac{2}{\sqrt{w_p}}$$

If, however, $w_p < w_c$ an infinite number of switching are required to reach the origin although the total time elapsed is still finite and sums to*

$$S(w_c, w_p) = \frac{2\sqrt{2}}{\sqrt{w_p}\sqrt{1 + \dfrac{w_p}{w_c}}} \,, \quad w_p < w_c$$

Finally, if $w_p > w_c$, the systems chatters to the origin along the curve $\dfrac{\dot{x}|\dot{x}|}{2w_c} + x = 0$. The time in this case is computed to be

---

* These results are derived in reference (6). However, the results presented here differ by a factor

$$\frac{1 + \sqrt{w_c^2 - w_p^2}}{1 - (w_o - w_p)}$$

from those of reference (6). The present calculations are believed to be correct.

$$S(w_c, w_p) = \frac{\sqrt{2}\sqrt{1 + \frac{w_p}{w_e}}}{\sqrt{w_p}}, \quad w_p > w_c$$

Since $S(w_c, w_p)$ is a monotonically decreasing function of w_p, it follows that

$$\max_{w_p} S(w_c, w_p) = S(w_c, .8)$$

For $w_p = .8$ the value of w_c which minimizes S, by the assumption of an optimal controller structure, is $w_c = .8$, therefore

$$\min_{w_c} \max_{w_p} S(w_c, w_p) = S(.8, .8) = 2.25$$

Now

$$\min_{w_c} S(w_c, w_p) = \min\left(\frac{\sqrt{2}\sqrt{1 + \frac{w_p}{.8}}}{\sqrt{w_p}}, \frac{2\sqrt{2}}{\sqrt{w_p}\sqrt{1 + \frac{w_p}{1.2}}}\right)$$

Furthermore for $.8 \leq w_p \leq .97$

$$\min_{w_c} S(w_c, w_p) = \frac{\sqrt{2}\sqrt{1 + \frac{w_p}{.8}}}{w_p}$$

and for $.8 \leq w_p \leq 1.2$

$$\min_{w_c} S(w_c, w_p) = \frac{2\sqrt{2}}{\sqrt{w_p}\sqrt{1 + \frac{w_p}{1.2}}}$$

The curve $\min_{w_c} S(w_c, w)$ may therefore be seen to be monotonically decreasing with w_p and hence maxmin of S occurs at $w_p = .8$, with

$$\max \min S(w_c, w_p) = S(.8, .8) = 2.25$$

From result (3) the pure strategy $w_p = .8$ and $w_c = .8$ is optimal and the controller design parameter is $w_c = .8$. For this choice of w_c

the time required to reach the origin is guaranteed not to exceed 2.25 seconds for $.8 \leq w_p \leq 1.2$.

5. ACKNOWLEDGEMENTS

The authors are indebted to Professor L. Shaw for his discussions on various points in the theory of games and to Professor D. Youla for pointing out some of the limitations of the game-theoretic approach to sensitivity problems. The work reported herein was supported by the U.S. Army Research Office - Durham under Contract No. DA 31-124-ARO-D316 and by the Joint Services Electronics Program (U.S. Army, U.S. Navy and U.S. Air Force) under Contract No. AF 49(638)-1402.

6. REFERENCES

1. J. C. C. McKinsey, "Introduction to the Theory of Games", McGraw-Hill, New York, N. Y., 1952.

2. S. Karlin, "Mathematical Methods and Theory in Games, Programming, and Economics, Vol. II," Addison-Wesley, Reading, Mass., 1959.

3. R. F. Drenick, P. Dorato, "Optimality, Insensitivity and Game Theory", International Symposium on Sensitivity Analysis, Dubrovnik, Yugoslavia, September, 1964.

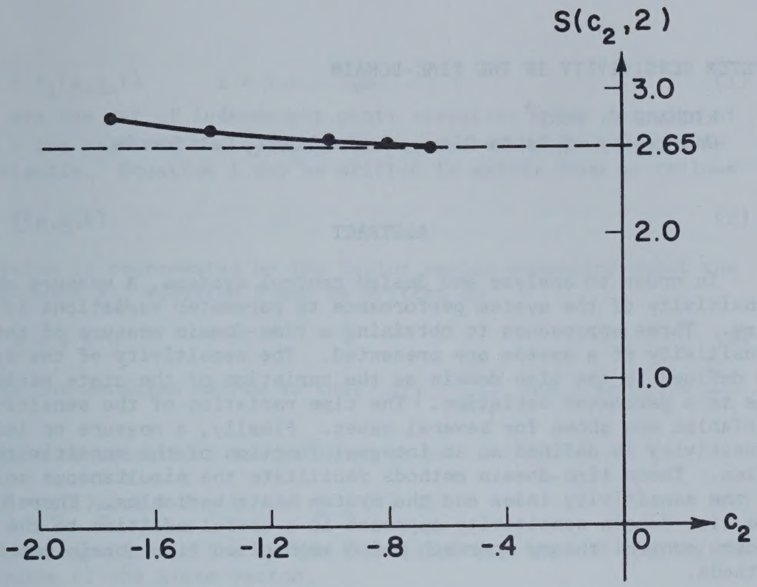
4. R. A. Rohrer, M. Sobral, Jr., "Sensitivity on Optimal System Design", IEEE Trans. on Automatic Control, Vol. AC-10, No. 1, pp. 43-48, January, 1965.

5. A. Kestenbaum, "Optimal Control Sensitivity Design", Master of Science Report, E. E. Dept., Polytechnic Institute of Brooklyn, June, 1965.

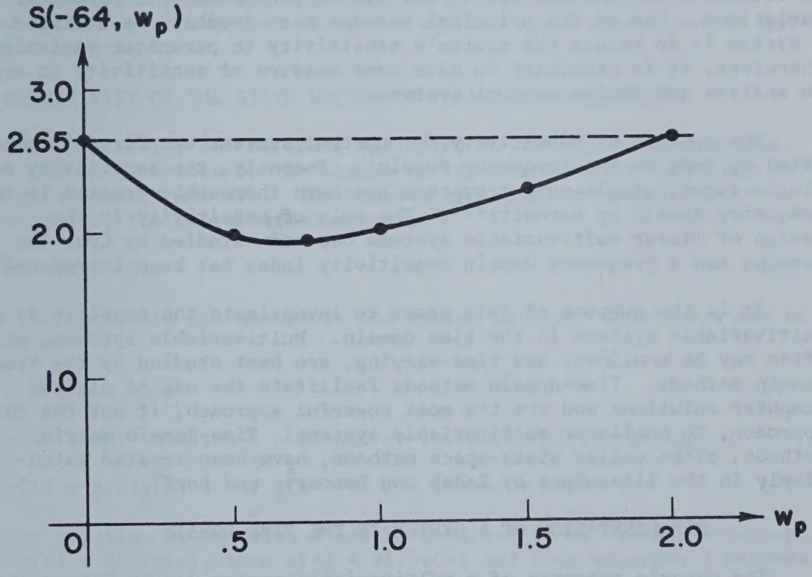
6. P. Shivdasani, "Sensitivity of Some Low Order Bang-Bang Control Systems", Master of Science Report, E. E. Dept., Polytechnic Institute of Brooklyn, June, 1965.

7. P. Dorato, "On Sensitivity in Optimal Control Systems", IEEE Trans. on Automatic Control, Vol. AC-8, No. 3, July, 1963.

8. M. Gavrilovic, R. Petrovic, and D. Siljak, "Adjoint Method in the Sensitivity Analysis of Optimal Systems", Jour. Franklin Institute, Vol. 276, No. 1, pp. 26-38, July, 1963.



(a) Graph of $S(c_2, 2)$ vs c_2



(b) Graph of $S(-.64, w_p)$ vs. w_p

Fig. 1 Graphs of $S(c_2, w_p)$

SYSTEM SENSITIVITY IN THE TIME-DOMAIN

RICHARD C. DORF*
University of Santa Clara, Santa Clara, California

ABSTRACT

In order to analyze and design control systems, a measure of the sensitivity of the system performance to parameter variations is necessary. Three approaches to obtaining a time-domain measure of the sensitivity of a system are presented. The sensitivity of the system is defined in the time domain as the variation of the state variables due to a parameter variation. The time variation of the sensitivity variables are shown for several cases. Finally, a measure or index of sensitivity is defined as an integral function of the sensitivity variables. These time-domain methods facilitate the simultaneous solution of the sensitivity index and the system state variables. Therefore, the time-domain sensitivity approach is a useful addition to the modern control theory approach which emphasizes time-domain matrix methods.

INTRODUCTION

For the analysis and design of control systems, it is imperative to determine the sensitivity of the system performance to parameter variations. One of the principal reasons that feedback is utilized in a system is to reduce the system's sensitivity to parameter variations. Therefore, it is necessary to have some measure of sensitivity in order to analyze and design control systems.

The concept of sensitivity for control systems was first investigated by Bode in the frequency domain[1]. Recently, the sensitivity of single-input, single-output systems has been thoroughly treated in the frequency domain by Horowitz[2,3]. The role of sensitivity in the design of linear multivariable systems has been studied by Cruz and Perkins and a frequency domain sensitivity index has been introduced[4].

It is the purpose of this paper to investigate the sensitivity of multivariable systems in the time domain. Multivariable systems, which often may be nonlinear and time-varying, are best studied by the time-domain methods. Time-domain methods facilitate the use of digital computer solutions and are the most powerful approach, if not the only approach, to nonlinear multivariable systems. Time-domain matrix methods, often called state-space methods, have been treated extensively in the literature by Zadeh and Desoer[5], and Dorf[6].

REPRESENTATION OF A SYSTEM IN THE TIME-DOMAIN

The dynamic response of a multivariable system may be represented by a set of first-order differential equations as follows

---

*R.C. Dorf is with the Department of Electrical Engineering, The University of Santa Clara, Santa Clara, California

$$\frac{dx_i}{dt} = f_i(\underline{x},\underline{u},t) \qquad i = 1,2,\ldots,n \tag{1}$$

where x_i are the set of independent state variables which describe at any time t the state of the system. The u_j are the signal inputs or forcing signals. Equation 1 may be written in matrix form as follows

$$\dot{\underline{x}} = \underline{F}(\underline{x},\underline{u},t). \tag{2}$$

If the system is represented by the Taylor series expansion about the operating point $\underline{x}_o$, equation 2 may be written as

$$\dot{\underline{x}} = \underline{A}\ \underline{x} + \underline{B}\ \underline{u} \tag{3}$$

where $\underline{A}$ is the Jacobian matrix so that $a_{ij} = \frac{\partial f_i}{\partial x_j}$ and B is $b_{ij} = \frac{\partial f_i}{\partial u_j}$.

The solution of equation 3 can be shown to be

$$\underline{x}(t) = \underline{\phi}(t)\underline{x}(o) + \int_o^t \underline{\phi}(t - \tau)\ \underline{B}\ \underline{u}\ (\tau)\ d\tau \tag{4}$$

where $\underline{\phi}(t) = \exp\ (\underline{A}t)$ is the system transition matrix, and $\underline{x}(o)$ is the initial value of the state vector.

When the nonlinearities in a system are of a large magnitude, such as for a relay control system, it is often impossible to use the linearized vector equation, equation 3. In that case, the original equation describing the system, equation 2, must be utilized. Nonlinear systems may be solved using the variable-gain approximation and computer solution^6.

SENSITIVITY OF THE STATE VECTOR TO PARAMETER VARIATIONS

The sensitivity of the state variables to variations of the parameters can be described by the change of the state vector $\Delta\underline{x}(t)$ due to a change of the parameters. The problem can be considered to be represented by the perturbation equation^7

$$\dot{\underline{x}} = (\underline{A} + \varepsilon\underline{D})\underline{x} \tag{5}$$

where $\varepsilon\underline{D}$ represents the change in the system matrix. Equation 5 may be written as

$$\dot{\underline{x}} = \underline{A}\ \underline{x} + \varepsilon\underline{D}\ \underline{x}. \tag{6}$$

The solution of equation 6 obtained by the perturbation method is

$$\underline{x}(t) = \underline{\phi}(t)\underline{x}(o) + \varepsilon \int_o^t \underline{\phi}(t - \tau)\ \underline{D}\ \underline{x}\ (\tau)\ d\tau + 0(\varepsilon^2) \tag{7}$$

The solution of equation 7 may be obtained in an iterative manner. Since $\underline{x}(t) = \underline{\phi}(t)\underline{x}(o)$, then $\underline{x}(\tau) = \phi(\tau)\underline{x}(o)$ and thus equation 7 becomes

$$\underline{x}(t) = \underline{\phi}(t)\underline{x}(o) + \varepsilon \int_o^t \underline{\phi}(t - \tau)\ \underline{D}\ \underline{\phi}(\tau)\underline{x}(o)\ d\tau \tag{8}$$

Therefore, the change in the state vector due to the change in the system matrix is

$$\Delta\underline{x}(t) = \varepsilon \int_o^t \underline{\phi}(t - \tau)\ \underline{D}\ \underline{\phi}(\tau)\underline{x}(o)\ d\tau. \tag{9}$$

For a stable system, the elements of $\phi(t)$ are decaying exponentials and thus equation 9 converges to zero as time increases without limit.

If the variation of the system parameters is equal for each coefficient of the system matrix, equation 5 becomes

$$\dot{\underline{x}} = (\underline{A} + \varepsilon\underline{A})\underline{x}. \tag{10}$$

Then the change in the state vector is

$$\Delta\underline{x}(t) = \varepsilon \int_o^t \underline{\phi}(t - \tau)\ \underline{A}\ \underline{\phi}(\tau)\underline{x}(o)\ d\tau. \tag{11}$$

Furthermore, $\underline{A}$ may be written as

$$\underline{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 0 & 0 & 0 & & 1 \\ -a_{n1} & -a_{n2} & -a_{n3} & \cdots & -a_{nn} \end{bmatrix}$$

and the system may be initially displaced from equilibrium so that

$$\underline{x}^T(o) = [1,0,0,\ldots,0].$$

Then equation 11 becomes

$$\Delta\underline{x}(t) = \varepsilon \int_o^t \underline{\phi}(t - \tau) \begin{bmatrix} \phi_{21}(\tau) \\ \phi_{31}(\tau) \\ \cdot \\ \cdot \\ \cdot \\ \phi_{n-1,1}(\tau) \\ \psi(\tau) \end{bmatrix} d\tau \tag{12}$$

where $\psi(\tau) = -a_{n1}\phi_{11} - a_{n2}\ \phi_{21} \cdots -a_{n,n-1}\phi_{n-1,1} - a_{nn}\phi_{n1}$.

Therefore, the change in the state vector can be evaluated for a change in the system parameter matrix by utilizing equation 11 or equation 12. In order to effectively illustrate this approach, consider the following example.

Example: Consider the system with the transfer function

$$G(s) = \frac{1}{(s+1)(s+2)} \quad \text{and the system matrix } \underline{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

The transition matrix for this system is

$$\underline{\phi}(t) = \begin{bmatrix} (2e^{-t}-e^{-2t}) & (e^{-t}-e^{-2t}) \\ (-2^{-t}+2e^{-2t}) & (-e^{-t}+2e^{-2t}) \end{bmatrix}. \tag{13}$$

When $\underline{x}^T(o) = [1,0]$, we obtain from equation 12

$$\Delta\underline{x}(t) = \varepsilon \int_o^t \underline{\phi}(t-\tau) \begin{bmatrix} \phi_{21}(\tau) \\ -2\phi_{11} \; -3\phi_{21} \end{bmatrix} d\tau$$

or, for example,

$$\Delta x_1(t) = \varepsilon \int_o^t \left[\phi_{11}(t-\tau)\; \phi_{21}(\tau) + \phi_{12}(t-\tau)(-2\phi_{11}(\tau) - 3\phi_{21}(\tau))\right] d\tau$$

$$\varepsilon \int_o^t \left[-2\phi_{11}(t-\tau)\; \phi_{12}(\tau) -2\phi_{12}(t-\tau)\; \phi_{11}(\tau) + 6\phi_{12}(t-\tau)\; \phi_{12}(\tau)\right] d\tau$$

$$= \varepsilon\left[e^{-t}(6-2t) - e^{-2t}(6+10t)\right].$$

The normalized change in $x_1(t)$ is given in Table I and is shown in Fig. 1.

THE SENSITIVITY OF THE CANONICAL STATE VARIABLES TO PARAMETER VARIATIONS

Alternatively, the sensitivity of the canonical state variables can be related to the parameter changes. The system differential equation in canonical form is

$$\dot{\underline{y}} = \underline{\Lambda}\, \underline{y} \tag{14}$$

where Λ is a diagonal matrix whose elements are the characteristic roots of the system when the roots are distinct. The diagonal matrix is obtained by utilizing a transformation matrix $\underline{T}$ so that $\underline{\Lambda} = \underline{T}^{-1}\underline{A}\,\underline{T}$. The solution of equation 14 is

$$\underline{y}(t) = e^{\underline{\Lambda} t}$$

or

$$y_i(t) = e^{\lambda i t}. \tag{15}$$

If the parameter variations can be related to a small change in the canonical system matrix $\underline{\Lambda}$, then we may write

$$\dot{\underline{y}} = (\underline{\Lambda} + \varepsilon\, \underline{\Lambda})\, \underline{y}. \tag{16}$$

Following the procedure of the preceding paragraphs, the solution of equation 16 is found to be

$$\underline{y}(t) = e^{\underline{\Lambda} t}(\underline{I} + \varepsilon\, \underline{\Lambda}\, t) \tag{17}$$

where $\underline{y}(o) = \underline{I}$. Then the change in a canonical state variable is

$$\Delta y_i = \varepsilon \lambda_i t e^{\lambda_i t}. \tag{18}$$

The maximum change in the canonical state variable is

$$\text{max. } \Delta y_i = -\varepsilon(.3678)$$

and occurs when $t = -1/\lambda_i$

THE SENSITIVITY DESCRIBED BY PARTIAL DERIVATIVES

The sensitivity of the system state vector can be described in terms of the variations of the parameters by considering the partial derivatives of the state variables with respect to each parameter. This is a particularly applicable approach for simulation of the system on an analog computer or digital computer[8].

The differential equation describing the system is

$$x_i = f_i(\underline{x},\underline{u},t) \qquad i = 1,2,\ldots,n. \tag{19}$$

Equation 19 may be rewritten as

$$g_i(\dot{\underline{x}},\underline{x},\underline{u},t) \tag{20}$$

or in matrix form as

$$\underline{g}(\dot{\underline{x}},\underline{x},\underline{u},t) = 0. \tag{21}$$

Now it is desired to obtain the variation in each state variable with respect to the parameter p_k; that is $\partial x_i/\partial p_k$. The partial derivative of equation 20 with respect to the parameter is

$$\frac{\partial g_i}{\partial \dot{x}_j}\frac{\partial \dot{x}_j}{\partial p_k} + \sum_{i,j=1}^{n,m} \frac{\partial g_i}{\partial x_j} \times \frac{\partial x_j}{\partial p_k} + \frac{\partial g_i}{\partial p_k} = 0 \tag{22}$$

where we assume u is independent of p_k. Since $\dfrac{\partial g_i}{\partial \dot{x}_i} = 1$ and $\dfrac{\partial g_i}{\partial \dot{x}_j} = 0$

when $j \neq i$, we obtain the following equation

$$\frac{\partial \dot{x}_j}{\partial p_k} + \sum_{i,j}^{n,m} \frac{\partial g_i}{\partial x_j} \times \frac{\partial x_j}{\partial p_k} + \frac{\partial g_i}{\partial p_k} = 0\ . \tag{23}$$

In matrix form we have

$$\frac{\partial \dot{\underline{\underline{x}}}}{\partial p_k} + \underline{\nabla}\ \underline{g} \times \frac{\partial \underline{\underline{x}}}{\partial p_k} + \frac{\partial \underline{g}}{\partial p_k} = 0 \tag{24}$$

or

$$\begin{bmatrix} \frac{\partial \dot{x}_i}{\partial p_k} \\ \frac{\partial \dot{x}_2}{\partial p_k} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} + \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial p_k} \\ \frac{\partial x_2}{\partial p_k} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} + \begin{bmatrix} \frac{\partial g_1}{\partial p_k} \\ \frac{\partial g_2}{\partial p_k} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = 0$$

The gradient of $\underline{g}$ with respect to $\underline{x}$ is the Jacobian matrix $\underline{J}$ and $\frac{\partial \underline{x}}{\partial p_k}$ will be defined as the vector $\underline{v}$. Then equation 24 may be written as

$$\dot{\underline{v}} = \underline{J}\ \underline{v} + \underline{z} \tag{25}$$

where $\underline{z} = - \frac{\partial \underline{g}}{\partial p_k}$. Equation 25 is of the same form as the linearized system equation, equation 3, obtained from equation 1 and is rewritten for comparison as

$$\begin{aligned} \dot{\underline{x}} &= \underline{J}\ \underline{x} + \underline{B}\ \underline{u} \\ &= \underline{J}\ \underline{x} + \underline{q}. \end{aligned} \tag{26}$$

The solution of equation 25 is

$$\underline{v}(t) = \underline{\phi}(t)\ \underline{v}(o) + \int_o^t \underline{\phi}(t - \tau)\ \underline{z}(\tau)\ d\tau \tag{27}$$

where $\phi(t) = \exp(-Jt)$. Therefore, we have the variation in the state vector due to a change in a simple parameter as

$$\frac{\Delta \underline{x}(t)}{\Delta p_k t} = \underline{v}(t). \tag{28}$$

The forcing function $\underline{z}(\tau)$ in the integrand of equation 27 relates the solution of the parameter sensitivity vector $\underline{v}$ to the original system state vector, $\underline{x}(t)$. If the parameter appears only in one term of the $\underline{g}$ matrix, for example, g_{mn}, then $\underline{z}(\tau) = z(\tau) = qx_n(t)$ where q is a constant. In general, $\underline{z}$ is comprised of several of the system state variables $x_i, x_j, \ldots, x_m$. In any case, the explicit solution of equation 27 requires the availability of several, if not all, of the state variables. Therefore, the simultaneous solution of equations 26 and 27 can be obtained by computer methods[8].

It would be desirable to know the variation in the state vector as each parameter p_k, $k = 1,2,3,\ldots,m$, is varied. Therefore, the solution of equation 25 could be repeated for each parameter. Alternatively, since $\underline{z} = \partial \underline{g}/\partial p_k$ is a function of the state variables, the solution of $\underline{v}$ for each parameter can be obtained simultaneously by utilizing parallel programs for the computer solution.

Now let us obtain the resulting change in each state variable for a linear system with parameter variations. The system matrix will be

$$A=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_{n1} & -a_{n2} & -a_{n3} & \cdots & -a_{nn} \end{bmatrix}$$

Then, for example, if the parameter that is varying is $-a_{n1} = p_1$, we have $\underline{z}^T = [0,0,\ldots,x_1]$. If the parameter varies a small amount so that we have $-a_{n1}(1+\varepsilon)$, and $\underline{v}(o) = 0$, we obtain from equation 27

$$\underline{v}(t) = \int_o^t \underline{\phi}(t-\tau)\, \underline{z}(\tau)\, d\tau = \int_o^t \begin{bmatrix} \phi_{1n}(t-\tau) \\ \phi_{2n}(t-\tau) \\ \vdots \\ \phi_{nn}(t-\tau) \end{bmatrix} x_1(\tau)\, d\tau. \qquad (29)$$

However, if $\underline{x}^T(o) = [1,0,0,\ldots,0]$, then $x_1(\tau) = \phi_{11}(\tau)$, and therefore

$$v_i(t) = \int_o^t \phi_{in}(t-\tau)\, \phi_{11}(\tau)\, d\tau. \qquad (30)$$

In order to effectively illustrate this approach, let us consider the following example.

Example: Reconsider the linear system of the previous example, that is,

$$\underline{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}.$$

The transition matrix is given in equation 13. Assume that the parameter $a_{21} = p_1$ increases by a small amount so that $a_{21} = -2(1+\varepsilon)$ and that $\underline{v}(o) = 0$. Then we obtain

$$\underline{v}(t) = \int_o^t \underline{\phi}(t-\tau)\, \underline{z}(\tau)\, d\tau \qquad (31)$$

where $\underline{z}(\tau) = \begin{bmatrix} 0 \\ x_1(\tau) \end{bmatrix}$. Therefore, we find that

$$\begin{aligned} v_1(t) &= \frac{\partial x_1}{\partial p_1} = \int_o^t \phi_{12}(t-\tau)\, \phi_{11}(\tau)\, d\tau \\ v_2(t) &= \frac{\partial x_2}{\partial p_1} = \int_o^t \phi_{22}(t-\tau)\, \phi_{11}(\tau)\, d\tau \end{aligned} \qquad (32)$$

when $\underline{x}^T(o) = [1,0]$ and thus $x_1(\tau) = \phi_{11}(\tau)$.

Evaluating equation 32 we obtain

$$\begin{aligned} v_1(t) &= e^{-t}(2t-3) + e^{-2t}(t+3) \\ v_2(t) &= e^{-t}(5-2t) - e^{-2t}(5+2t). \end{aligned} \qquad (33)$$

The time response of the sensitivity variables, $v_1(t)$ and $v_2(t)$, is given in Table II and is shown in Fig. 2.

It would be desirable to compare the accuracy of the sensitivity variables with the exact change of the state variables due to a change in a_{21} so that $a_{21} = -2(1+\varepsilon)$. The actual change in x (t) was calculated for $\varepsilon = .05$ by obtaining $x_1(t)$ when $a_{21} = -2$ and $a_{21} = -2.1$. Several values of the actual $\Delta x_1(t)$ are shown in Fig. 3. Since $\Delta x_1/\Delta p_1$ we have

$$\Delta x_1 = v_1(t)\ \Delta p_1 = v_1(t)(+\varepsilon\ a_{21}) = -.1\ v_1(t). \tag{34}$$

The curve of the change in $x_1(t)$ obtained from the sensitivity coefficient is shown in Fig. 3. The agreement is very good for this 5% increment in the parameter.

SENSITIVITY ANALYSIS UTILIZING THE TIME-DOMAIN MATRIX

The sensitivity of a system may be obtained in the time domain by the use of time-domain matrices. The time-domain matrix is a matrix expression of the convolution summation for a sampled system[6]. If the system is a continuous response system, the introduction of a sampler and hold circuit allows the use of the time-domain matrix method. A unity feedback control system is shown in Fig. 4. The output of the system is

$$\underline{c} = \underline{G}\ \underline{e} = \underline{G}\{\underline{I} + \underline{G}\}^{-1}\ \underline{r} \tag{35}$$

where $\underline{G}$ is the system matrix[6]. A system transfer matrix is defined as $\underline{T} = \underline{G}\{\underline{I} + \underline{G}\}^{-1}$ which relates the output to the system input. The sensitivity of the system is defined in the z-transform domain as

$$S_G^T(z) = \frac{\partial T(z)/T(z)}{\partial G(z)/G(z)} = \frac{1}{1 + G(z)} \tag{36}$$

where $T(z) = \dfrac{G(z)}{1 + G(z)}$. In an analogous manner, the time-domain sensitivity is defined as

$$\underline{S} = \{\underline{I} + \underline{G}\}^{-1} \tag{37}$$

This sensitivity function may be readily obtained as a time response and utilized for analysis and design purposes. In order to illustrate the utilization of this method, let us consider the following example.

Example: Consider the system shown in Fig. 4 when $G(s) = \dfrac{1}{s(s+1)}$ and the sampling period is $T = 1$ second. Then we obtain

$$\underline{G} = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ .3679 & 0 & 0 & 0 & \\ .7675 & .3679 & 0 & 0 & \\ .9145 & .7675 & .3679 & 0 & \\ .9685 & .9145 & .7675 & .3679 & \cdots \\ .9884 & .9685 & .9145 & .7675 & \cdots \\ \vdots & & & & \end{bmatrix}$$

The inverse of a lower triangular matrix is readily obtained, yielding for the first column[9] $\underline{S_1} = [1, -.3679, -.632, -.400. +.0001, +.253, \ldots]^T$. The time-domain sensitivity response is shown in Fig. 5. The area under the absolute value of $\underline{S_1}(t)$ may be considered as a measure of the system sensitivity.

In order to compare the sensitivity of an uncompensated and a compensated system, consider the digital compensation system shown in Fig. 6. The system transfer matrix for this system is

$$\underline{T} = \underline{G}\,\underline{D}\,\{\underline{I} + \underline{G}\,\underline{D}\}^{-1} \qquad (38)$$

and the sensitivity is

$$\underline{S} = \{\underline{I} + \underline{G}\,\underline{D}\}^{-1} \,. \qquad (39)$$

A digital compensator which will reduce the overshoot for a step input to one-half the uncompensated value while reducing the rise time by 25% is $D(z) = \frac{1.565(z - .561)(z + .456)(z - 1)}{z^3}$. In this case, the time-domain sensitivity is obtained utilizing equation 38 and is plotted in Fig. 5. Using the function

$$I = \int_0^{\infty} |S_1(t)|\,dt \qquad (40)$$

as the criterion of sensitivity performance, the compensated system results in a sensitivity reduction of approximately 25%.

A SENSITIVITY INDEX

In order to compare one system or design with another, some numerical measure of a system's sensitivity is desired. In general, a useful time-domain index is obtained from in integral function of the sensitivity time response. Therefore, as noted in the preceding section, a useful index is

$$I = \int_0^{t_1} |\underline{S}_1(t)|\,dt \qquad (41)$$

where $\underline{S}_1(t)$ is the sensitivity response obtained from the time-domain matrix equation. Alternatively, any even function may be used for the integral measure. For example, the quadratic function

$$I = \int_0^{t_1} (\underline{S}_1(t)^2)\,dt \qquad (42)$$

might be utilized.

The sensitivity coefficients discussed earlier would provide a measure of the sensitivity performance by utilizing the equation

$$I = \int_0^{t_1} g(\underline{v}(t))\,dt. \qquad (43)$$

A useful index for computer evaluation would be

$$I = \int_0^{t_1} |\underline{v}(t)|\,dt. \qquad (44)$$

This type of performance index would be particularly useful for system parameter optimization by iterative computer methods.

For the sensitivity evaluation by perturbation methods developed earlier, one might use the sensitivity index

$$I = \int_0^{t_1} g(\Delta\underline{x}(\quad))\,dt.$$

The sensitivity indices are readily evaluated and provide a reasonable measure of the sensitivity of the system to the parameters of interest.

CONCLUSIONS

In order to analyze and design control systems, a measure of the sensitivity of the system performance to parameter variations is necessary. Three approaches to obtaining a time-domain measure of the sensitivity of a system have been presented. These time-domain methods facilitate the simultaneous computer solution of the sensitivity index and the system state variables. Therefore, the time-domain sensitivity approach is a useful addition to the modern control theory approach which emphasizes time-domain matrix methods.

REFERENCES

1. H.W. Bode, "Network Analysis and Feedback Amplifier Design", Van Nostrand, New York, 1945.

2. I.M. Horowitz, "Fundamental Theory of Automatic Linear Feedback Control Systems", IRE Trans. on Automatic Control, v AC-4, Dec. 1959; pp 5-19.

3. I.M. Horowitz, "Synthesis of Feedback Systems", Academic Press, New York, 1963.

4. J.B. Cruz, Jr., and W.R. Perkins, "The Role of Sensitivity in the Design of Multivariable Linear Systems", Proc. of the Nat'l Elec. Conf., 1964; pp 742-745.

5. L.A. Zadeh and C. Desoer, "Linear System Theory", McGraw-Hill, New York, 1963.

6. R.C. Dorf, "Time Domain Analysis and Design of Control Systems", Addison-Wesley, Reading, Mass., 1965.

7. R. Bellman, "Perturbation Techniques in Mathematics, Physics, and Engineering", Holt, Rinehart and Winston, Inc., New York, 1964.

8. H.F. Meissinger, "The Use of Parameter Influence Coefficients in Computer Analysis of Dynamic Systems", Proc. of the Western Jt. Computer Conf., 1960; pp 181-192.

9. R.A. Frazer, W.J. Duncan, A.R. Collar, "Elementary Matrices", Cambridge Univ. Press, 1952.

ACKNOWLEDGMENT

The author is indebted to his colleague, Dr. D.D. Siljak, for many fruitful discussions, and especially acknowledge his suggestion of integral sensitivity indices.

TABLE I

$\frac{\Delta x_1}{\varepsilon}$	0	-1.-14	-.694	-.206	-.030	-.001
t	0	.5	1.0	2.0	5.0	10

TABLE II

t	0	.5	1.0	2.0	5.0	10.0
$v_1(t)$	0	.075	.173	.227	.048	.001
$v_2(t)$	0	.219	.157	-.030	-.035	-.001

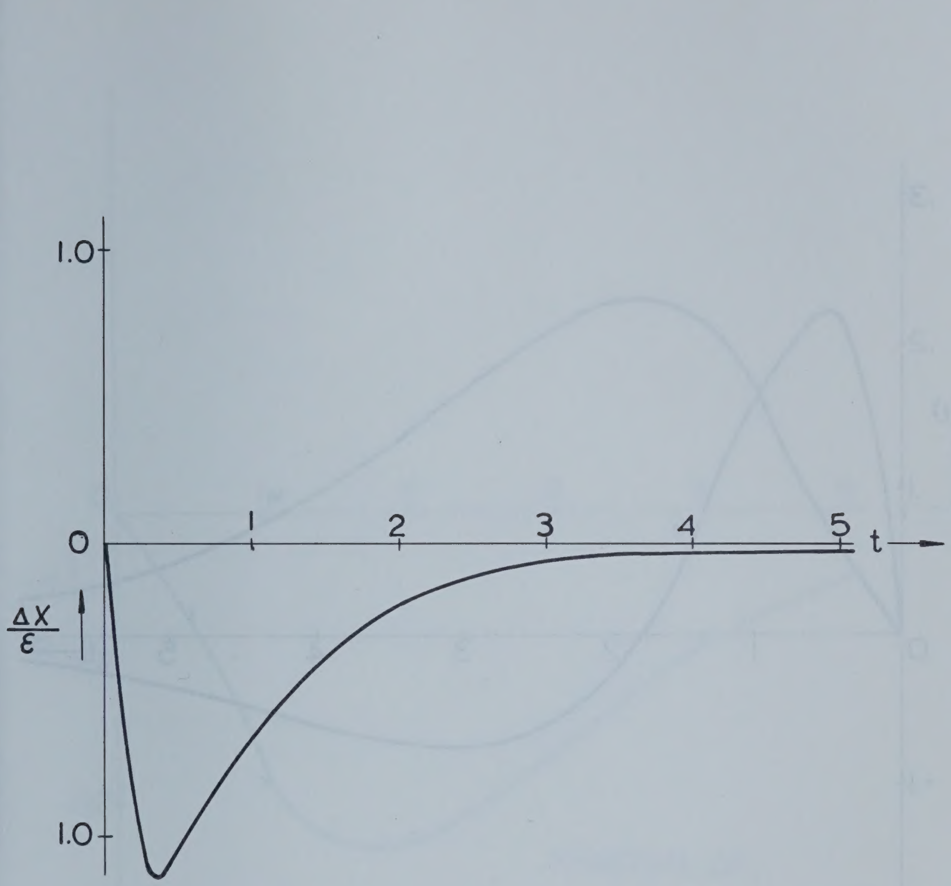


FIG. 1

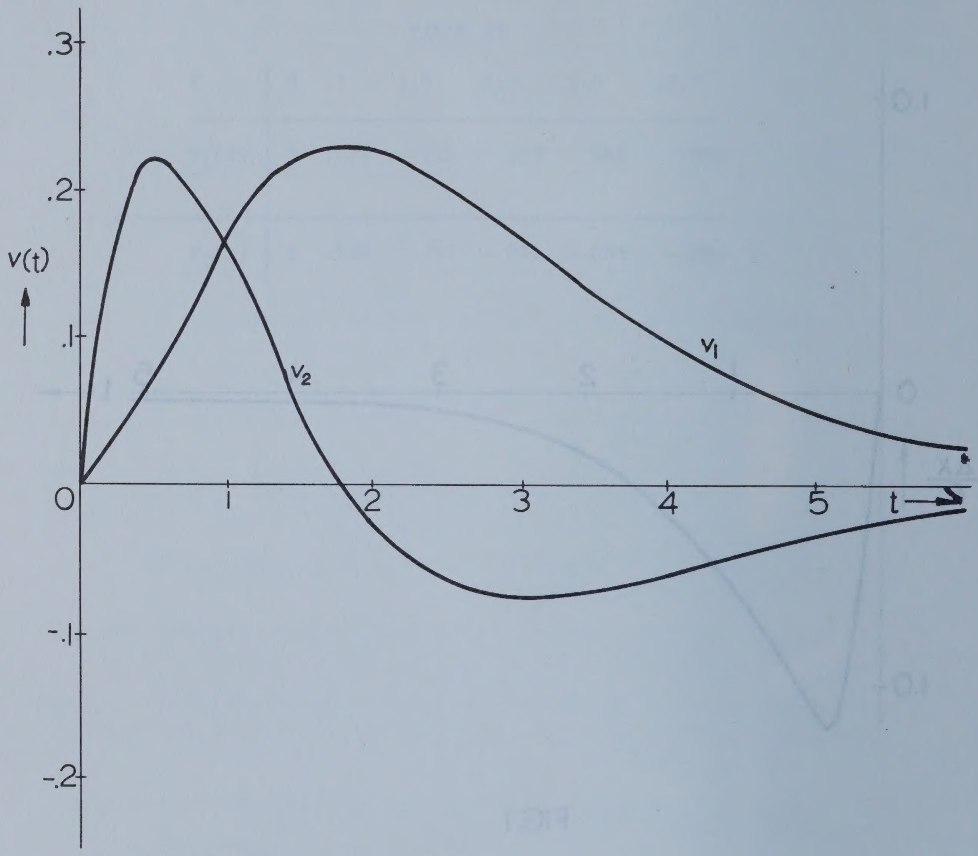


FIG. 2

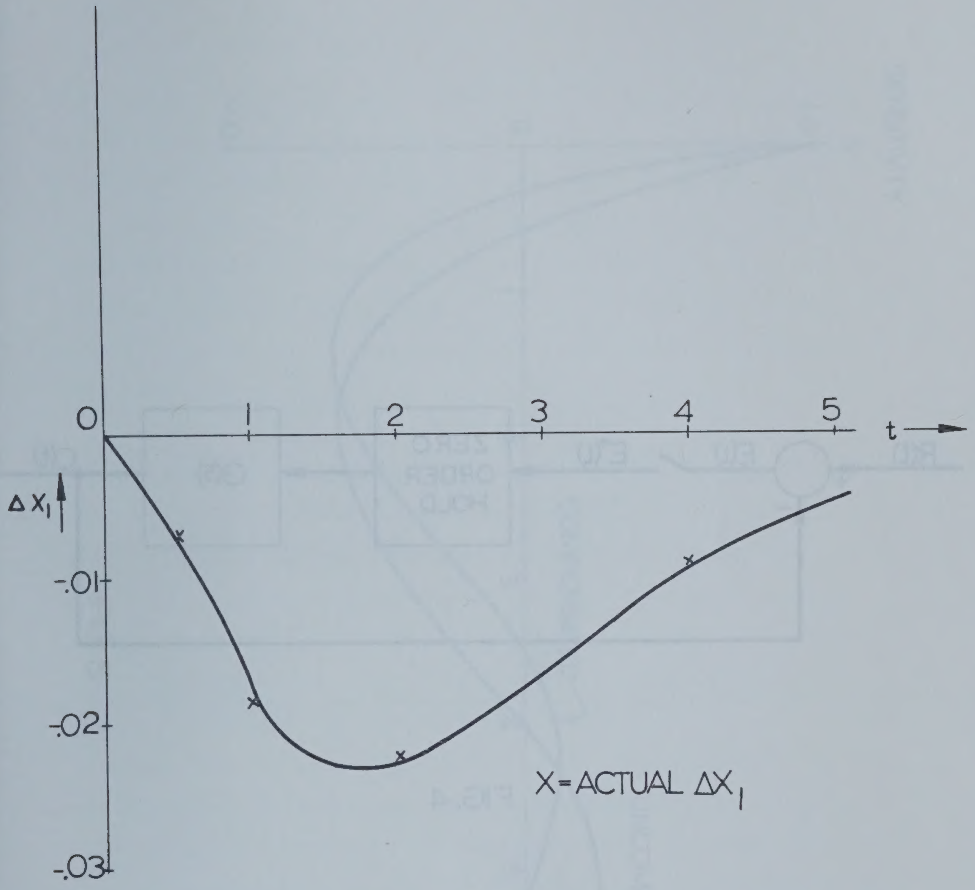


FIG.3

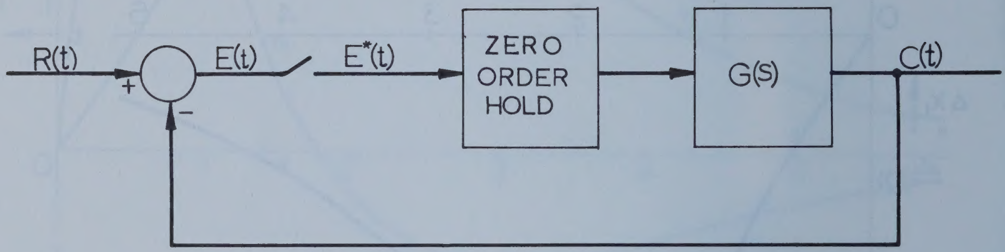

R(t)
+
−
E(t)
E*(t)
ZERO ORDER HOLD
G(S)
C(t)

FIG. 4

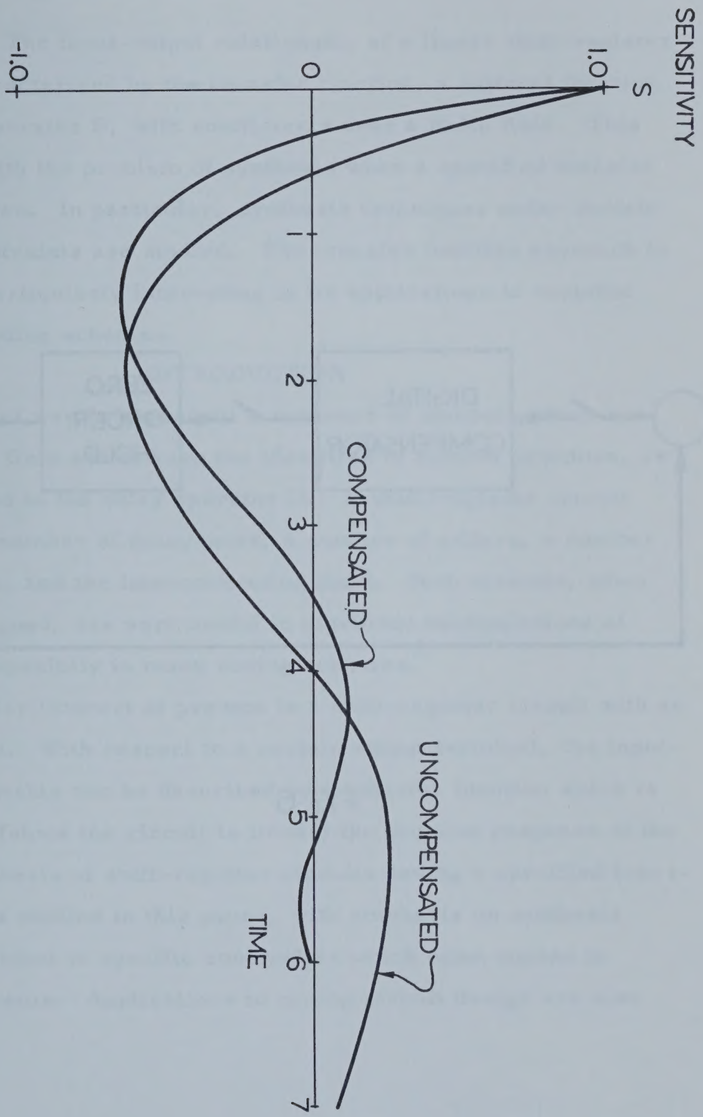

SENSITIVITY
S
1.0
0
-1.0
1
2
3
4
5
6
7
TIME
COMPENSATED
UNCOMPENSATED

FIG. 5

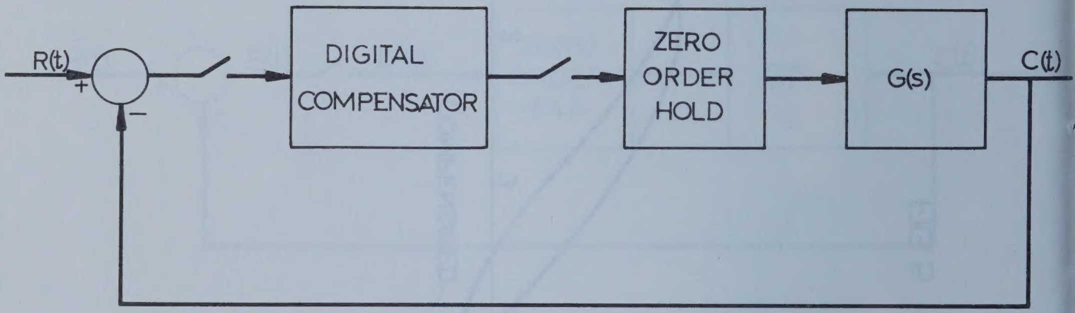

R(t)
+
−
DIGITAL
COMPENSATOR
ZERO
ORDER
HOLD
G(s)
C(t)

FIG. 6

TRANSFER FUNCTION SYNTHESIS OF LINEAR SHIFT-REGISTER CIRCUITS

D. T. Tang
IBM Watson Research Center
Yorktown Heights, New York

ABSTRACT: The input-output relationship of a linear shift-register circuit is characterized by the transfer function, a rational function in the delay operator D, with coefficients over a finite field. This paper deals with the problem of synthesis when a specified transfer function is given. In particular, synthesis techniques under certain system's constraints are studied. The transfer function approach is shown to be particularly interesting in its applications to variable redundancy coding schemes.

INTRODUCTION

A convenient way to represent a sequence of symbols which are elements of a field and to keep the identities of symbol positions, is to use a series in the delay operator D.[1] A shift-register circuit consists of a number of delay units, a number of adders, a number of multipliers, and the interconnecting links. Such circuits, when properly designed, are very useful in algebraic manipulations of sequences, especially in many coding schemes.[2]

Of particular interest at present is a shift-register circuit with an input terminal. With respect to a certain output terminal, the input-output relationship can be described by a transfer function which is equivalent to (since the circuit is linear) the impulse response of the circuit. Synthesis of shift-register circuits having a specified transfer function is studied in this paper, with emphasis on synthesis techniques subject to specific constraints which often appear in practical systems. Applications to coding circuit design are also shown.

SPECIFICATION OF THE TRANSFER FUNCTION

The transfer function to be realized can often be written directly from the knowledge of the desired algebraic operation to be performed on the input sequence. For example, if the transfer function is just a polynomial f(D), then the input series in D will be multiplied by the polynomial f(D) at the output; or if the transfer function is 1/f(D), the inverse of a polynomial, then the input will be divided by f(D). In general, the transfer function of a finite shift-register circuit is a rational function $T(D) = f(D) / g(D)$, where f(D) and g(D) are polynomials in D. The rational function expression of a transfer function can be obtained from the series expression of the impulse response.

MULTIPLICATION AND DIVISION

While the standard multiplication and division circuits are well known,[1, 2] it is of some interest to observe that the circuit shown in Fig. 1 can either multiply(with switch position at 1) or divide (with switch position at 2) the input by $f(D) = D^3 + D^2 + 1$. This is generally true even in a nonbinary case, which is shown in Fig. 2.

It is easy to check and see that when switches are at position 1, the transfer function is (output/inout) = f(D); but when switches are at position 2, we have (output/input = 1/f(D). It should be noted that f(D) is assumed to have the constant term equal to 1.*

We shall see that the identification of variables at terminal points with transfer functions equal to $T_i(D) - 1$ as well as $T_i(D)$ often proves to be very helpful in understanding the circuit performance and in optimizing the circuit design in general.

INPUT-FREE SOLUTIONS

Let's consider a division circuit such as the one shown in Fig. 2 with switches set at position 2. The impulse response of the circuit is 1/f(D). Since $f(D) \mid (1 - D^n)$, where n is the period of f(D)[2, 3], we

---

*This is equivalent to the multiplication as divisions of the input polynomial by a monic polynomial g(x) with higher order coefficients sent first. In such a case $f(D) = g^*(D)$, where $g^*(D)$ is the reciprocal polynomial of g(D).

may write $f(D) \cdot h(D) = 1 - D^n$, or

$$\frac{1}{f(D)} = \frac{h(D)}{1-D^n} = h(D)\,(1 + D^n + D^{2n} + \ldots). \qquad (1)$$

Therefore the impulse response of the division circuit is just coefficients of h(D) repeated with a period of n.

Similarly, if the initial contents of registers in the above division circuit are not all zeroes, the output will not be zero even if there is no input. From Fig. 2, one can readily write the condition for such a solution: $s\,[\,f(D) - 1\,] = -s$, or $s \cdot f(D) = 0$. Since s, the input-free solution, must also assume the same period n, if we write

$s = s'\,(\ldots + D^{-2n} + D^{-n} + 1 + D^n + D^{2n} + \ldots)$, we must have

$$s' \cdot f(D) \equiv 0 \quad \mathrm{mod}\,(1-D^n), \qquad (2)$$

$$\text{or} \quad s' \equiv 0 \quad \text{mode } h\,(D), \qquad (3)$$

which states that the solution of an input-free circuit whose characteristic polynomial is f(D) must be coefficients of a certain multiple of h(D) repeated with the period n.

It should be clear that, in order to have a shift-register circuit function properly as characterized by its transfer function, the initial contents of the registers must be zero. If this is not the case, a certain close-loop solution or a transient response (when the transfer function is a polynomial, indicating that no feed-back connection is present) will be added to the output.

FACTORIZATION OF f(D) -1 AND f(D)

If f(D) - 1 can be factored, then it is a simple matter to connect realizations of these factors in series and from which to obtain a multiplication circuit, a division circuit, or a maximum length sequence generator, each having a characterizing polynomial equal to f(D). In fact, this technique can be used to obtain a circuit with a minimum number of adders and constant multipliers. Applications of this technique to binary cases have been studied in Ref. 4.

In many applications, such as in certain coding schemes, f(D) itself may be the product of two or more polynomials. For simplicity, let $f(D) = f_1(D) \cdot f_2(D)$. Figure 3 shows a terminated bridged T circuit including two components with transfer functions equal to $f_1(D) - 1$ and $f_2(D) - 1$ respectively. The overall transfer function of this circuit is

$$T(D) = [f_1(D) - 1][f_2(D) - 1] + [f_1(D) - 1] + [f_2(D) - 1]$$

$$= f_1(D) \cdot f_2(D) - 1 = f(D) - 1. \qquad (4)$$

The above technique can be repeated if further factorization of $f_1(D)$ or $f_2(D)$ is possible. This technique can either be used in an effort to minimize the total number of adders and constant multipliers, or be used in a system with multiple modes of operation, such as a variable-redundancy coding system[5]. Note that if we disconnect the part with transfer function $f_1(D) - 1$ in Fig. 3, the overall transfer function becomes $f_2(D) - 1$, and vice-versa.

CONSTRAINED SYNTHESIS OF POLYNOMIALS

Both synthesis techniques discussed in the last section are special forms of the following constrained synthesis problem: To realize an arbitrary polynomial p(D) with the total number of registers equal to $|p(D)|$, the degree of p(D), and with the realization of another q(D) as a component. In particular, we would like to have the configuration as shown in Fig. 4.

It is clear that for the total number of registers to be $|p(D)|$, we must have $|p(D)| \geq q(D)|$. Suppose that it is possible to write

$$p(D) = q(D) \cdot a(D) + r(D). \qquad (5)$$

with $|r(D)| \leq |a(D)|$, then realizations of p(D) with the confuguration shown in Fig. 4 exist. One of such realizations is shown in Fig. 5.

Actually one may first connect any realization of a(D) in series with a realization of q(D), and then make up the difference r(D) by proper tapping and weighting. This is possible since the impulse responses at various points in the realization of a(D) clearly contain a set of basis which will generate any polynomial of degree no more than

$|a(D)|$. If Eq. (5) cannot be written with $|r(D)| \leq |a(D)|$, realization of p(D) is possible with the number of registers equal to $|p(D)|$ only when tappings from inside the realization of q(D) are allowed.

If we let $p(D) = f_1(D) \cdot f_2(D) - 1$ and $q(D) = f_2(D) - 1$ in Eq. (5), we have

$$f_1(D) \cdot f_2(D) - 1 = [f_2(D) - 1]\, f_1(D) + [f_1(D) - 1], \tag{6}$$

corresponding to the circuit shown in Fig. 3.

SYNTHESIS OF RATIONAL FUNCTIONS

A transfer function which takes the general form of the following rational function

$$T(D) = \frac{f_o + f_1 D + f_2 D^2 + \ldots + f_n D^n}{1 + g_1 D + g_2 D^2 + \ldots + g_m D^m} = \frac{f(D)}{g(D)} \tag{7}$$

is always realizable with the total number of registers no more than max (m, n). A standard realization of Eq. (7) is shown in Fig. 6.

Since it is again possible to first realize $1/g(D)$ in any fashion and next obtain the factor f(D) by proper tapping and weighting (note that some extra registers are necessary if $n > m$), synthesis techniques by factoring either g(D) or g(D) - 1 can be used. A circuit with a minimum number of adders and multipliers can often be obtained by selecting one of many possible factorizations. Algebraic manipulations of Eq. (7) which simplify g(D) by factorization techniques and in the meanwhile express f(D) with as many factors appearing in the various levels of factorization procedures of g(D) as possible usually lead to simple realizations. This is demonstrated in the following example.

Example 1. Consider the following rational function over the binary field

$$T(D) = \frac{D^5 + D^4 + D^3 + D^2 + 1}{D^9 + D^8 + D^7 + D^6 + D^2 + D + 1}$$

The standard realization technique leads to a circuit with twelve adders. However, the same function can be written as:

$$T(D) = \frac{1 + D^2 (D + 1)(D^2 + 1)}{1 + D(D + 1)[1 + D^5 (D^2 + 1)]}$$

which gives a much simpler realization as shown in Fig. 7.

SOME CODING APPLICATIONS

Rather than studying the shift-register contents only, it is helpful (and often sufficient) to analyze the time sequences (series in D) at some key points in a circuit by means of the corresponding transfer functions. This approach has played an important role in the hardware design of some variable redundancy coding schemes[5, 6].

The circuit shown in Fig. 8 is a decoder for a binary coding scheme capable of operating in the following two modes:[5]

Mode 1: $g_1(x) = x^4 + x + 1$ is the generator polynomial of a (15, 11) single error correcting code.

Mode 2: $g(x) = g_1(x)\, g_2(x) = (x^2 + x + 1)(x^4 + x + 1)$ is the generator polynomial of a (15, 9) burst-3 correcting code.

The validity of the decoding circuitry can be easily seen by recognizing the use of the circuit of Fig. 3 which gives $(y/z) = g_1^*(D) + 1$ when switches M are at position 1, and $(y/z) = g_1^*(D)\, g_2^*(D) + 1$ when switches M are at position 2. The remaining part of this circuit is the same as that used in Refs. 2 and 7, with errors corrected sequentially. Switch N is at position 1 during the first n bits, and is at position 2 during the next n bits.

Figure 9 shows an interesting dual circuit which is equivalent to the one shown in Fig. 8. Here two division circuits corresponding to $g_1(x) = x^2 + x + 1$ and $g_2(x) = x^4 + x + 1$ are connected in series.

To show the equivalence of these two decoding circuits, we first identify the transfer functions obtained at certain key points in the circuit shown in Fig. 9: For Mode 1 operation, we have

$$\frac{y}{x} = \frac{u}{v} = \frac{g_2^*(D) + 1}{g_2^*(D)} = \frac{D^4 + D^3}{D^4 + D^3 + 1} \qquad (8)$$

and

$$\frac{z}{x} = \frac{z}{v} = \frac{1}{g_2^*(D)} = \frac{1}{D^4 + D^3 + 1}$$

And for Mode 2 operation, we have

$$\frac{y}{x} = \frac{w+u}{x} = \frac{w}{x} + \frac{v}{x}\,\frac{u}{v} = \frac{g_1^*(D)+1}{g_1^*(D)} + \frac{g_2^*(D)+1}{g_1^*(D)\,g_2^*(D)}$$

$$= \frac{g_1^*(D)\,g_2^*(D)+1}{g_1^*(D)\,g_2^*(D)} = \frac{D^6+D^5+D^4+D^3}{D^6+D^5+D^4+D^3+1}$$

and

$$\frac{z}{x} = \frac{v}{x}\cdot\frac{z}{v} = \frac{1}{g_1^*(D)\,g_2^*(D)} = \frac{1}{D^6+D^5+D^4+D^3+1} \tag{9}$$

It takes little effort to show that both Eqs. (8) and (9) also hold for the corresponding input-output pairs in Fig. 8. These circuits are therefore equivalent as far as their decoding function is concerned.

Similar technique finds it spplication in a situation where Reed-Solomon code is used in a storage device. The circuit in Fig. 10 is capable of operating in two modes: With switches M at position 1, the circuit can be used as a separable code encoder. (Here switches K are at position 1 for the duration of information symbols and are at position 2 for the duration of check symbols). However, with switches at position 2, it becomes a part of the decoder where syndromes are obtained.

It is easy to see that syndromes are readily obtained at the syndrome outputs when switches M are at position 2. With switches M at position 1, these $(x-\alpha^i)$ dividers are connected in series, and (z/x) ,$g^*(D) = (1-\alpha D)(1-\alpha^2 D)\ldots(1-\alpha^m D)$. If information symbols represented by $I(x)$ are coded as $C(x) = x^m I(x) + r(x)$ where $r(x)$ is the residue of $x^m I(x)$ modulo $g(x)$, then the last m of n symbols of z must be zero if $x = C^*(D)$, since $g(x) \mid C(x)$. But $x + y = z$, if follows that $x = C^*(D)$ implies $x = -y = D^{n-m} r^*(D)$ for the last m of n symbols. Thus, with switches K at position 2, the above situation is observed, and the check symbols $D^{n-m} r^*(D)$ are obtained by negating y as shown in Fig. 10.

CONCLUSIONS

Various synthesis techniques of shift-register circuits have been presented in this paper. They have been shown to be useful in circuit minimization as well as in circuit designs of certain coding schemes.

REFERENCES

1. Huffman, D. A., "The Synthesis of Linear Sequential Coding Networks," Information Theory, Colin Cherry (Ed), Academic Press, New York, 1956.

2. Peterson, W.W., "Error Correcting Codes," MIT Technology Press, Cambridge, Mass. and Wiley, New York, New York, 1961.

3. Elspas, B., "The Theory of Autonomous Linear Sequential Networks," IRE Transactions, CT-6, pp 45-60, March, 1959.

4. Roth, H.H., "Linear Binary Shift Register Circuits Utilizing a Minimum Number of Mod-2 Adders," IEEE Transactions, IT-11, No. 2, pp 215-220, April, 1965.

5. Tang, D. T., and Chien, R. T., "Cyclic Product Codes and Their Implementation," to appear in Information and Control.

6. Tang, D. T., "Dual Codes as Variable Redundancy Codes," IEEE International Convention Record, pp. 220-226, March, 1965.

7. Meggitt, J. E., "Error-Correcting Codes and Their Implementation for Data Transmission Systems," IRE Transactions, IT-7, pp. 234-244, October, 1961.

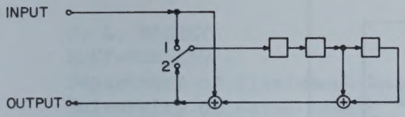

INPUT
1
2
OUTPUT

Fig. 1

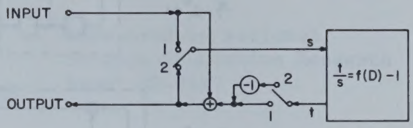

INPUT
1
2
s
OUTPUT
-1
2
1
t
$\frac{t}{s} = f(D) - 1$

Fig. 2

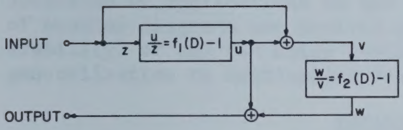

INPUT
z
$\frac{u}{z} = f_1(D) - 1$
u
v
$\frac{w}{v} = f_2(D) - 1$
OUTPUT
w

Fig. 3

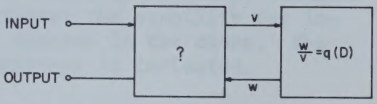

INPUT
?
v
$\frac{w}{v} = q(D)$
OUTPUT
w

Fig. 4

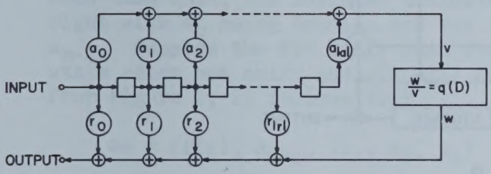

a_0
a_1
a_2
$a_{|a|}$
v
INPUT
$\frac{w}{v} = q(D)$
w
r_0
r_1
r_2
$r_{|r|}$
OUTPUT

Fig. 5

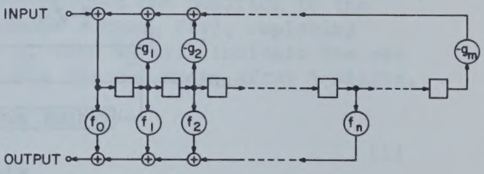

INPUT
$-g_1$
$-g_2$
$-g_m$
f_0
f_1
f_2
f_n
OUTPUT

Fig. 6

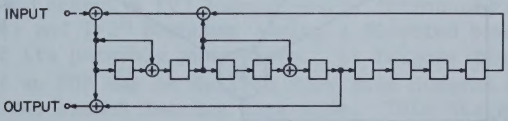

INPUT
OUTPUT

Fig. 7

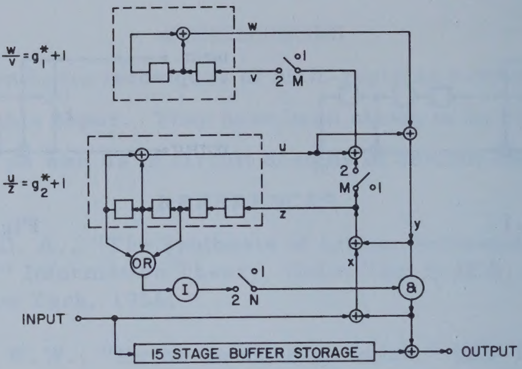

w
w/v = g₁* + 1
v
1
2 M
u
u/z = g₂* + 1
2
M
1
z
y
OR
x
I
1
2 N
&
INPUT
15 STAGE BUFFER STORAGE
OUTPUT

Fig. 8

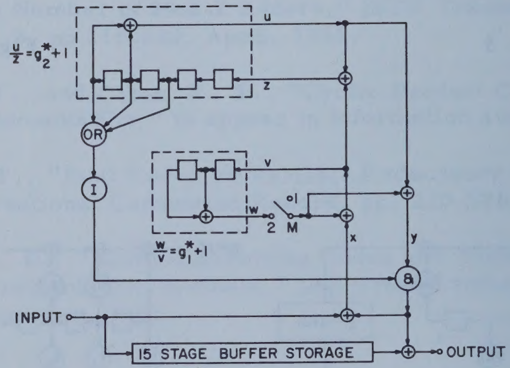

u
u/z = g₂* + 1
z
OR
I
v
1
w
2 M
w/v = g₁* + 1
x
y
&
INPUT
15 STAGE BUFFER STORAGE
OUTPUT

Fig. 9

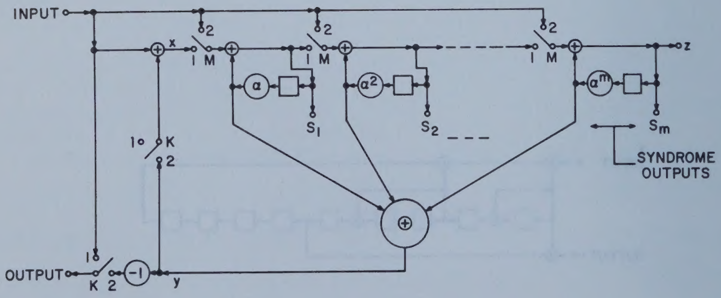

INPUT
2
2
2
x
1 M
1 M
1 M
z
α
α²
αᵐ
S₁
S₂
Sₘ
SYNDROME OUTPUTS
1
K
2
1
OUTPUT
K 2
-1
y

Fig. 10

A NEW DIAGRAM FOR FEEDBACK SHIFT-REGISTERS

J. L. MASSEY
RUEY-WEN LIU
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

Sponsored by National
Science Foundation Research
Grant GP-2547

ABSTRACT

A new and easily constructed diagram for feedback shift registers (FSR's) is developed which contains all the information on the traditional Good's diagram. The new diagram has the advantage that the diagram for m-stage FSR's is a sub-diagram of that for (m+1)-stage FSR's. The usefulness of this new diagram, called an annular diagram, is illustrated by applications to the theory of stability of FSR's. Classes of annular diagrams are derived which guarantee the stability (or instability) of any FSR which satisfies any diagram in the class. The generalization to multi-point feedback registers is indicated.

INTRODUCTION

In this paper, a new diagram for the study of feedback shift-registers, called an annular diagram, is introduced. The basic properties of this diagram are derived and its usefulness is illustrated by applications to the theory of stability of feedback shift-registers.

A general autonomous binary feedback shift-register, or FSR, is shown in Figure 1. The binary m-tuple $\underline{s} = (s_{m-1}, s_{m-2}, \ldots, s_1, s_o)$ is the state of the FSR where each s_j is the output of a unit delay stage. Each time unit, the register contents are shifted one position to the right with s_o being dropped and the feedback output, $f(\underline{s})$, replacing s_{m-1}. Suppose the FSR is in some state $\underline{s}$, then $\theta\underline{s}$ will indicate the new state after one shift and $\theta^j\underline{s}$ will indicate the new state after j shifts. From Figure 1, it follows that

$$\theta\underline{s} = (f(\underline{s}), s_{m-1}, \ldots, s_2, s_1) \tag{1}$$

so that the successor of $\underline{s}$ is either $(0, s_{m-1}, \ldots, s_2, s_1)$ or $(1, s_{m-1}, \ldots, s_2, s_1)$ depending on whether $f(\underline{s}) = 0$ or $f(\underline{s}) = 1$. There are 2^{2^m} distinct m-stage FSR's corresponding to the 2^{2^m} distinct choices of $f(\underline{s})$ as a boolean function of m variables.

The state diagram of an FSR is a directed linear graph of 2^m nodes and 2^m branches, each node corresponding to a state of the FSR with each state connected to its successor by a directed branch. In 1946, I. G. Good [1] and N. G. deBruijn [2] independently introduced a diagram with 2^m nodes (states) and $2 \cdot 2^m$ branches having a directed branch from each state to both of its possible successors. It follows that every possible state diagram of an FSR can be derived from this diagram by some deletion of one and only one branch leaving each node. This diagram has been commonly referred to as Good's Diagram in the FSR literature and has been extensively exploited [3,4,5]. Good's diagram for m=3 is shown in Figure 2.

For some purposes, the usefulness of Good's diagram is impaired by the facts that (1) the diagram is quite cumbersome except for very small values of m, and (2) the diagram of order m+1 does not contain that of order m as a sub-diagram. These features of Good's diagram are overcome in the annular diagram described in the next section without loss of any information contained in Good's diagram.

DESCRIPTION OF THE ANNULAR DIAGRAM

To facilitate the description of the annular diagram, we first arithmetize the operation of FSR's. With the state $\underline{s} = (s_{m-1}, s_{m-2}, \ldots s_1, s_o)$, we identify the integer S given by

$$S = s_{m-1}2^{m-1} + s_{m-2}2^{m-2} + \ldots + s_1 2 + s_o \ , \qquad (2)$$

and we let θS and $\theta^j S$ denote the integers identified with $\theta \underline{s}$ and $\theta^j \underline{s}$ respectively. We also identify $F(S)$ and $f(\underline{s})$, i.e., $F(S) = f(\underline{s})$ when S is the integer identified with $\underline{s}$. With the new notation, Eq. (1) becomes

$$\theta S = [S/2] + F(S) \cdot 2^{m-1} \qquad (3)$$

where [] denotes the integer part of the enclosed number.

The annular diagram is a convenient graphical presentation of Eq. (3). The diagram for m=5 is shown in Figure 3. The diagram is formed by placing the integers (states) $0,1,2,\ldots,2^m-1$ into m+1 annuli containing 1, 1, 2, 4, ..., 2^{m-1} states respectively. Beginning at the center, the states are arranged in increasing order proceeding clockwise around each annulus. Consider any state S. If $F(S) = 0$ then from Eq. (3),

$$\theta S = [S/2]$$

and θS is always the next state radically inward on the diagram (state 0 being counted inward to itself.) State [S/2] will be called the drop successor of S. On the hand, when $F(S) = 1$, from Eq. (3)

$$\theta S = [S/2] + 2^{m-1} \ .$$

The next state $[S/2] + 2^{m-1}$ will be called the jump successor of S. The number 2^{m-1} will be called the jumping number and is always the smallest integer on the outermost annulus of the annular diagram. The jump successor of any state is simply its drop successor plus the jumping number. Thus, the two possible successors of any state may be read easily from the annular diagram. For example, from Figure 3, the possible successors of state 29 are its drop successor, state 14 (the next state radially inward), or its jump successor state 30 (the drop successor plus the jumping number). Thus, all the information on Good's diagram is available on the annular diagram, and the latter has the features both (1) that it is easily constructed and (2) the diagram for m+1 contains that for m as a sub-diagram.

The complete description of a particular FSR is simply a designation for each state S whether its actual successor is its drop successor or its jump successor. If the former, S is called a drop state; if the latter, S is called a jump state. The specification will be made on the annular diagram by making each drop state with a "-" superscript and each jump state with a "+" superscript. For example, the m=4 annular diagram of Figure 4 is marked according to the feedback function $f(\underline{s}) = s_3 + s_o$ where + denotes "modulo-two" addition. From Figure 4 we see that state 0 is its own successor and that the other 15 states lie on a single closed cycle, viz.

$$15^- \rightarrow 7^+ \rightarrow 11^- \rightarrow 5^+ \rightarrow 10^+ \rightarrow$$

$$13^- \rightarrow 6^- \rightarrow 3^+ \rightarrow 9^- \rightarrow 4^- \rightarrow$$

$$2^- \rightarrow 1^+ \rightarrow 8^+ \rightarrow 12^+ \rightarrow 14^+ \rightarrow$$

SOME PROPERTIES OF THE ANNULAR DIAGRAM

Let $A_j (j = 0, 1, \ldots, m)$ denote the set of states on the jth annulus proceeding outward from the origin as shown in Figure 5(a). Then $A_o = \{0\}$, $A_1 = \{1\}$, $A_2 = \{2,3\}$. In general, for j>0,

$$A_j = \{n+2^{j-1} \mid n = 0, 1, \ldots 2^{j-1}-1\} \quad .$$

Let A_j^{**} denote the set of jump successors of the states in A_j, and let A_j^* denote the set of drop successors. From the preceeding section, we have immediately

Property 1:

$$A_j^* = \begin{array}{l} A_{j-1}, \ j = 1, 2, \ldots, m \\ A_o, \ j = 0 \end{array}$$

and similarly

Property 2:

$$A_j^{**} = \begin{array}{l} \{n + 2^{j-2} + 2^{m-1} \mid n = 0, 1, \ldots 2^{j-2}-1\} \quad , \ j = 2,3,\ldots, m \\ \{2^{m-1}\} \ , \ j = 0, 1 \end{array}$$

Since the jumping number 2^{m-1} is the smallest integer on the outermost annulus, it follows that all jump successors lie on the outermost annulus whereas all drop successors lie on inner annuli. The sets A_j^{**} and A_j^* have the interesting geometric locations on the annular diagram shown in Figure 5(b). This structure will be exploited in the next section.

Figure 5(b) provides insight into the structure of allowable state-sequences in FSR's. Consider some FSR in state S, then θS lies either on the next inner annulus or on the outermost annulus. Hence, the "operating point", ie., the location of the current state of the FSR, can move only in series of radially inward steps interspersed with a series of jumps on the outermost annulus.

Two other properties of the annular diagram will prove useful in the next section. Let S^* and S^{**} denote the drop successor and jump successor of S respectively.

Property 3:

$S^* \leq S$ with equality if and only if $S = 0$.

Property 4:

$S^{**} \geq S$ with equality if and only if $S = 2^m-1$.

Property 3 follows immediately from Figure 5(b) by noting that, except for $S = 0$, S^* lies on an annulus interior to S. To prove property 4, we note that

$$S^{**} = [S/2] + 2^{m-1}$$

and hence

$$S^{**} - S = 2^{m-1} - [(S+1)/2]. \qquad (4)$$

Since $S \leq 2^m-1$ for all states, the right hand side of (4) is positive except for $S = 2^m-1$.

APPLICATION TO STABILITY OF FSR'S.

A state such that $\theta S = S$ is said to be an equilibrium state of an FSR. From properties (3) and (4), it follows that states 0 and 2^m-1 are the only possible equilibrium states. 0 is an equilibrium state if $F(0)= 0$ and 2^m-1 is an equilibrium state if $F(2^m-1) = 1$. An FSR is said to be stable if 0 is an equilibrium state and for every state S there is some integer N such that $\theta^N S = 0$. From the point of view of the annular diagram, an FSR is stable if from any initial location the operating point eventually moves to the origin and remains there. Stability of FSR's has been shown to be an important consideration in the design of error-correcting circuits for use with convolutional codes [6]. In this section, we consider methods of determining stability, or instability, directly from examination of the annular diagram.

We adopt the convention of placing plus signs (+) or minus signs (-) in some region of an annular diagram to indicate a jump region or a drop region respectively. An FSR will be said to satisfy an annular diagram that has been so marked if every state lying wholly or partly in a jump region is a jump state, and in a drop region is a drop state. An annular diagram will be called a stable annular diagram if every FSR which satisfies it is stable, and an unstable annular diagram if no FSR which satisfies it is stable. By proving that some diagram is a stable (or unstable) annular diagram, one proves the stability (or instability) of a whole class of FSR's, namely the set of all FSR's which satisfy the diagram. If the diagram has M unspecified states, this class contains 2^M FSR's since F(S) may be chosen 0 or 1 arbitrarily on each of the M unspecified states.

Consider the annular diagram of Figure 6(a). We show next that any FSR which satisfies this diagram is stable. There are $M = 2^m-1$ unspecified states on the diagram since only one state on the outermost annulus (namely 2^m-1) is specified as a drop state. Proof that Figure 6(a) is stable then identifies a class of 2^{2^m-1} stable FSR's. To make the proof, we note first that any state in the drop region moves to state 0 in at most m steps. It remains to show that every unspecified state must eventually reach the drop region. If an unspecified state S is a drop state, then θS is already in the drop region. If S is a jump state then, by property 4, θS lies closer clockwise to state 2^m-1 than does S. In this case, there must be some j such that $\theta^j s$ satisfies $2^{m-1} \leq \theta^j s \leq 2^m-1$ and is a drop state.

Thus, every unspecified state eventually reaches the drop region which completes the proof.

Consider next the annular diagram of Figure 6(b), by the same argument as above, every unspecified state on the outermost annulus eventually reaches a state on A_{m-1}. But the successor of every state in A_{m-1} is in the drop region since the drop region includes both A_{m-1}^* and A_{m-1}^{**} (see Figure 5(b)). Hence this diagram is also a stable annular diagram. More generally, the same reasoning shows that

Theorem 1:

For any such j that $0 \leq j \leq m-1$, the diagram of Figure 6(c) is a stable annular diagram.

We note the power of this theorem in that it establishes the existence of a class of stable annular diagrams, each diagram of which guarantees the stability of an entire class of FSR's.

A second class of stable annular diagrams is given in Figure 7. Consider the diagram in Figure 7(a). Any state on A_{m-1} or A_m eventually reaches A_{m-2} according to the same reasoning used before. Any unspecified state in $A_1, A_2, \ldots A_{m-2}$ reaches the origin either via all drop state successors or via the fan-shaped drop region if one of its successors is a jump state. Hence the diagram is a stable annular diagram. Similar reasoning establishes:

Theorem 2:

For any j such that $0 \leq j \leq m-2$, the diagram of Figure 7(b) is a stable annular diagram.

As a final example of a class of stable annular diagrams, we consider the diagram of Figure 8 which has both drop regions and jump regions specified. $\emptyset$ is any angle less than 360°. Any state in the drop region moves to the origin in at most m steps. Any interior state either drops to the origin or else it or one of its successors is a jump state. The successor of the jump state must be on the outermost annulus. But, by property 4, any state in the jump region on the outermost annulus eventually reaches the drop region. Thus all states eventually reach the origin and remain there. We have thus shown:

Theorem 3:

For any angle $\emptyset < 360°$, the diagram of Figure 8 is a stable annular diagram.

The diagram in this class with the fewest specified states is obtained by taking $\emptyset$ so large that only state 2^m-1 on the outermost annulus lies in the drop region. In this case $M = 2^{m-1}-m$ states are unspecified. Thus

$2^{2^{m-1}-m}$ FSR's are guaranteed stable from this diagram.

Lastly, we give one example to illustrate how the reasoning of this section can be used to generate classes of unstable annular diagrams.

The simplest such class is given by
Theorem 4:

For any j such that $0 \leq j \leq m$, the diagram of Figure 9 is an unstable annular diagram.

The proof follows simply by noting that states can drop toward the origin in steps of only one annulus at a time. Thus for j>0 no state on A_j, A_{j+1}, ... A_m can reach the origin since annulus A_j (which is specified as all jump states) cannot be penetrated. If j = 0 state 0 is specified as a jump state. Since state 0 is then not an equilibrium state, no FSR which satisfies the diagram can be stable.

SUMMARY AND CONCLUSIONS

A new diagram for FSR's has been introduced and several properties of the diagram were derived. The diagram was then applied to the theory of stability of FSR's. The diagram permitted establishing by simple reasoning of the stability or instability of large classes of FSR's. The importance of the diagram is that stability or instability was ascertained without a detailed examination of the feedback functions of the corresponding FSR's.

All the results herein have been derived for feedback shift registers wherein the feedback output is returned only to the last stage of the register (FSR's.) The results can be applied to a more general case where the feedback is added modulo-two to an arbitrary set of stages of the register simply by using the known isomorphism between such registers and FSR's [7].

The authors found the annular diagram useful in several aspects of FSR behavior besides stability questions. It is hoped that this paper will stimulate research workers in the field to find further applications for this diagram.

REFERENCES

[1] I. G. Good, "Normal Recurring Decimals," Journal of London Mathematical Society, Vol. 21, Part 3, 1946, pp. 167-169.

[2] N. G. deBruijn, "A Combinatorial Problem," Proceedings Koninklyke Nederlandse Akademie van Wetenshappen, Vol. 49, Part 2, 1946, pp. 758-764.

[3] S. W. Golomb and L. R. Welch, "Nonlinear Shift-Registers Sequences," JPL Memorandum 20-149, Pasadena, Calif., Oct. 25, 1957.

[4] P. R. Bryant and R. D. Killick, "Nonlinear Feedback Shift Registers," IRE Trans. on Electronic Computers, EC-11, No. 3, June, 1962.

[5] K. B. Magleby, "The Synthesis of Nonlinear Feedback Shift Registers," Stanford Electronics Labs., Stanford, Calif., Tech. Rept. 6207-1, Oct. 1963.

[6] J. L. Massey & R. Liu, "Application of Lyapunov's Direct Method to the Error-Propagation Effect in Convolutional Codes," IEEE Trans.

on Information Theory, IT-10, No. 3, July, 1964, pp. 248-250.

[7] J. L. Massey & R. Liu, "Equivalence of Nonlinear Shift Registers," IEEE Trans. on Information Theory, IT-10, No. 4, Oct., 1964, pp. 378-379.

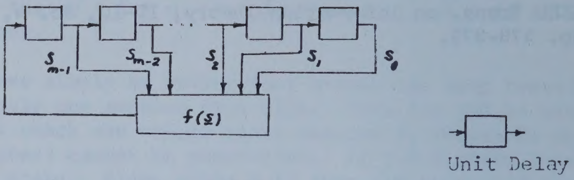

S_{m-1}
S_{m-2}
S_2
S_1
S_0
f(s)
Unit Delay

Fig. 1 General FSR

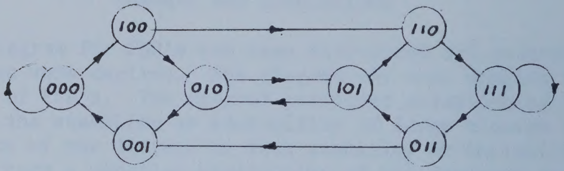

100
110
000
010
101
111
001
011

Fig. 2 Good's Diagram for m=3

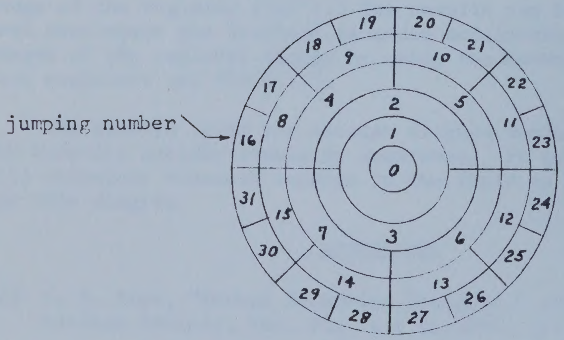

jumping number
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

Fig. 3 Annular Diagram for m=5

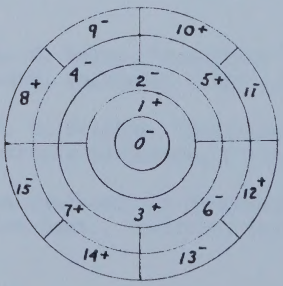
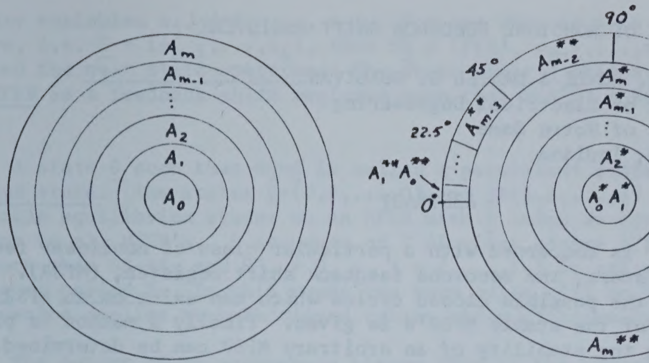

9-
10+
4-
2-
5+
8+
11-
1+
0-
15-
12+
7+
3+
6-
14+
13-

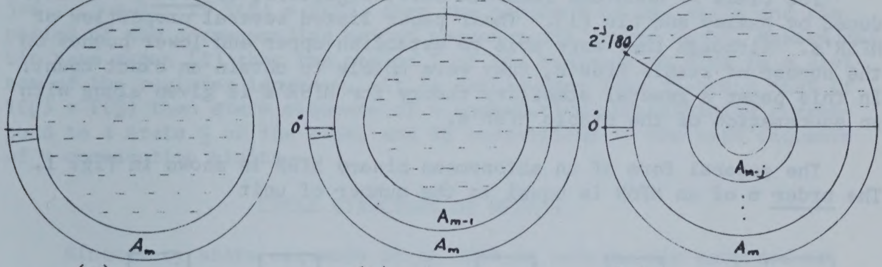
Fig. 4 Annular Diagram for m=4 marked for f($\underline{s}$) = $s_3 \oplus s_0$


A_m
A_{m-1}
⋮
A_2
A_1
A_0
90°
A_{m-2}**
45°
A_m*
A_{m-1}**
A_{m-3}**
22.5°
A_{m-1}*
A_1** A_0**
0°
A_2*
A_0* A_1*
A_m**

(a) Designation of Annuli

(b) Location of Jump Successors(**) & Drop Successors (*)

Figure 5


A_m
0°
A_{m-1}
A_m
2^{-j}·180°
0°
A_{m-j}
A_m

(a)

(b)

(c)

Figure 6 A Class of Stable Annular Diagrams

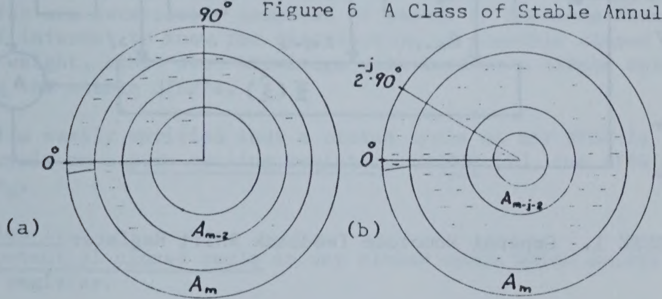

90°
0°
(a)
A_{m-2}
A_m
2^{-j}·90°
0°
A_{m-j-2}
(b)
A_m

Figure 7 Another Class of Stable Annular Diagrams

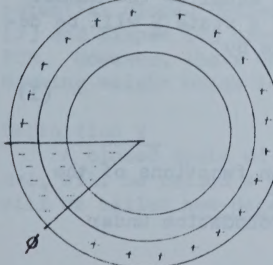


Fig. 8 A Class of Stable Annular Diagrams

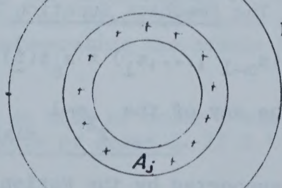


Fig. 9 A Class of Unstable Annular Diagrams

CYCLE STRUCTURE OF MONOTONE FEEDBACK SHIFT REGISTERS[1]

FREDERIC J. MOWLE & DANIEL D. SULLIVAN
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

ABSTRACT

This paper is concerned with a particular class of nonlinear feedback shift registers, the monotone feedback shift register, (MFSR). An enumeration of the possible closed cycles which can exist on an MFSR and an enumeration of the stable MFSR's is given. Finally a method is presented by which the stability of an arbitrary MFSR can be determined.

INTRODUCTION

The class of monotone feedback shift registers, or MFSR's was introduced by Massey and Liu [1]. Their paper listed several properties of MFSR's. Although they were able to establish upper and lower bounds on the number of stable MFSR's, they were unable to obtain an exact count. In this paper a general stability theory for MFSR's is given along with an enumeration of the stable MFSR's.

The general form of an autonomous binary MFSR is shown in Fig. 1. The order m of an MFSR is equal to the number of unit delay stages. The binary m-tuple $\underline{S} = (s_{m-1}, \dots, s_1, s_0)$ is the state of the MFSR, where each s_j is the output of one of the m unit delay stages. The Hamming weight of a binary m-tuple is equal to the number of its non-zero components. The Hamming weight of a state $\underline{S}$ will be denoted by $W(\underline{S})$. The feedback function $F(\underline{S})$ is given by

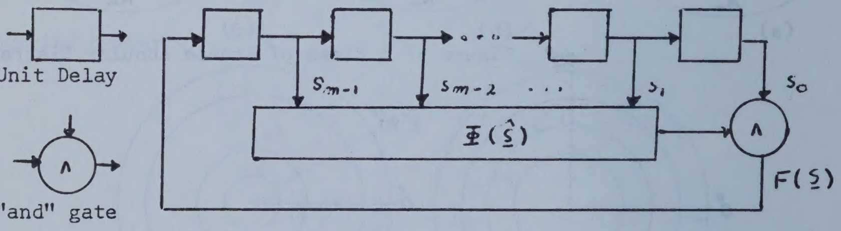


FIGURE 1. General Monotone Feedback Shift Register

$$F(\underline{S}) = s_0\Phi(s_{m-1}, \dots, s_1) \equiv s_0\Phi(\hat{\underline{S}}) \tag{1}$$

where $\Phi(\hat{\underline{S}})$ can be any of the $2^{2^{m-1}}$ distinct boolean functions of the binary variables s_j, j=1, ..., m-1. For any FSR, if $\underline{S}$ is the present state, i.e. $\underline{S} = (s_{m-1},\ldots,s_o)$, then $\theta\underline{S} = (F(\underline{S}), s_{m-1},\ldots,s_1)$. "$\theta$" is called the *next state operator*. The state after j shifts will be $\theta^j\underline{S}$. An *MFSR* is a feedback shift register such that $W(\theta^k\underline{S}) \leq W(\underline{S})$ for all $k\geq 1$.

---

[1] This work was supported by the National Science Foundation under Grant GP-2547.

A state $\underline{S}$ such that $\theta\underline{S}=\underline{S}$ is called a persistent state or an *equilibrium state*. The states $\underline{0}=(0,0,\ldots,0)$ and $\underline{1}=(1,1,\ldots,1)$ are the only possible equilibrium states on an MFSR with $\underline{0}$ being an equilibrim state common to all MFSR's. A sequence of T distinct states (T>1) $\underline{S}_1,\underline{S}_2,\ldots,\underline{S}_T$ such that $\theta\underline{S}_T=\underline{S}_1$ and $\theta\underline{S}_j = \underline{S}_{j+1}$, j = 1,...,T-1 is called a *closed cycle* of *period T*. In this paper the term cycle will be used when the discussion applies to both closed cycles and equilibrium states.

The output of an autonomous binary MFSR is the sequence of binary numbers: $Y(\underline{S})$, $Y(\theta\underline{S})$, $Y(\theta^2\underline{S}),\ldots$; where the *output function* $Y(\underline{S})$ can be any of the 2^{2^m} distinct boolean functions of the binary variables s_j, j=0,..., m-1. If $Y(\underline{S}) = F(\underline{S})$, then the period T of any cycle and the period of the corresponding output sequence are always equal. Massey and Liu have shown [1] that the period T of any cycle on an MFSR is a divisor of the register length m. It is easily seen from figure 1 that if $Y(\underline{S}) = F(\underline{S})$ then every sequence of m consecutive output binits correspond to a state $\underline{S}$ of the MFSR, and $\theta\underline{S}$ corresponds to the next sequence of m consecutive binits.

CYCLE STRUCTURE OF MFSR's

Since every state sequence on an MFSR is necessarily composed of states whose Hamming weight is monotonically non-increasing, all cycles on an MFSR are necessarily composed of states of the same Hamming weight. It is of interest to know the distribution of possible closed cycles by Hamming weight, since such knowledge will lead to a simple method of enumerating the stable MFSR's.

It is easily verified that a closed cycle on any MFSR is also a closed cycle on a *pure cycling register*, or *PCR* [2], the MFSR with $F(\underline{S}) = s_o$.

Definition 1.

A *potential closed cycle* is any closed cycle which exists on a pure cycling register.

Whether or not a potential closed cycle is in fact a closed cycle on a particular MFSR, depends on the feedback function $F(\underline{S})$. By determining the complete cycle structure of the pure cycling register, all potential closed cycles on an MFSR will be obtained. Golomb and Welch [2] and Elspas [3] have enumerating the total number of cycles for the PCR. However, their enumerations do not yield the cycle structure by Hamming weight which is useful in enumerating the stable MFSR's.

Definition 2

A closed cycle of period T on an MFSR of order m such that d = m/T, d>1, will be called a *degenerate cycle of order d*. If d = 1 the cycle will be called *non-degenerate*.

Since every closed cycle on a PCR of order m is a potential closed cycle on every MFSR of order m, a state $\underline{S}$ will be called degenerate of order d if it lies on a closed cycle of degeneracy d on the PCR. A degenerate state of order d will be of the form

$$\underline{S} = (\underbrace{a_1,a_2,\ldots,a_T}_{1}, \underbrace{a_1,\ldots,a_T}_{2},\ldots,\underbrace{a_1,\ldots,a_T}_{d}) \tag{2}$$

where $(a_1,a_2,\ldots,a_T)$ is a non-degenerate state (d=1) on an MFSR of order T, and the a_j's are binary numbers.

Lemma 1

A potential closed cycle of period T and degeneracy d=m/T consisting of states with Hamming weight w, $0<w<m$, exists on an MFSR if and only if d is an integer which divides both m and w.

Proof

Consider the MFSR defined by $F(\underline{S})=Y(\underline{S})=s_o$, i.e. the PCR. Starting in any state $\underline{S}$ having Hamming weight $W(\underline{S})=w$, $o<w<m$, a closed cycle will result. Let the resulting output sequence, which also will be cyclic, have period T and degeneracy d. Since the states on the closed cycle are of the form given by equation 2, there must be w/d ones and (m-w)/d zeros in the cyclic output sequence. If d does not divide both w and m, a fractional number of ones or zeros would result which is clearly impossible.

If d is an integer which divides both w and m, then

$$\underline{S} = (\underbrace{a_1,\ldots,a_T}_{1}, \underbrace{a_1,\ldots,a_T}_{2},\ldots,\underbrace{a_1,\ldots,a_T}_{d})$$

with

$a_j = 1, j = 1,\ldots,w/d$

$a_j = 0, j = w/d + 1,\ldots, T$

and m=dT, is a state which lies on a potential closed cycle of period T and degeneracy d. Q.E.D.

Having shown that for each integer d which divides both the register length m and the Hamming weight w, at least one potential closed cycle exists on an MFSR, the exact number of potential closed cycle of a given period remains to be determined.

Since m and w are both finite, only a finite number of distinct integers exist which are divisors of both m and w, $o<w<m$. Let D_w be the ordered set of all such divisors, i.e.

$$D_w = \{d_i \mid d_i \text{ divides both m and w}, d_{i+1} < d_i\} \tag{3}$$

The number of distinct states on an MFSR of order m which have Hamming weight w, $o<w<m$, is equal to the number of different ways of arranging w ones and m-w zeros in m positions. Letting N(m,w) denote this number, one obtains

$$N(m,w) = \binom{m}{w} = \frac{m!}{w!(m-w)!} \tag{4}$$

One might expect that the number of distinct states of Hamming weight w and degeneracy d_i on an MFSR of order m would be equal to $N(m/d_i, w/d_i)$. However, it is easily seen that some states would be counted more than once using this relation:

Example: the state (1,0,1,0,1,0,1,0) which has degeneracy four would be counted by $N(\frac{8}{4},\frac{4}{4})$ $N(\frac{8}{2},\frac{4}{2})$ and $N(\frac{8}{1},\frac{4}{1})$. Thus if d_j and $d_i \varepsilon D_\omega$, $i<j$ such that d_j divides d_i, then the states of degeneracy d_i will be counted again by $N(m/d_j, w/d_j)$. The following lemma is now evident.

Lemma 2

The number of distinct states of Hamming weight w, $o<w<m$ and degeneracy d_i on an MFSR of order m, denoted by $N^*(m,w,d_i)$, is given by the recursion relations:

$$N^*(m,w,d_1) = N(m/d_1, w/d_1)$$

$$N^*(m,w,d_i) = N(m/d_i, w/d_i) - \sum_{\substack{j=1 \\ d_i | d_j}}^{i-1} N^*(m,w,d_j) \qquad i > 1 \tag{5}$$

the summation being over only those d_j which d_i divides.

Theorem 1

There are exactly $Z_d(m,w,d_i)$ potential closed cycles of period $T_i = m/d_i$ and degeneracy d_i of Hamming weight w, $o<w<m$ on an MFSR of order m where

$$Z_d(m,w,d_i) = (d_i/m)N^*(m,w,d_i) \tag{6}$$

Proof

Let B be the set of all states of Hamming weight w and degeneracy d_i. From Lemma 2 there are $N^*(m,w,d_i)$ such states in B and from Lemma 1 $N^*(m,w,d_i)$ is at least equal to m/d_i. If a state $\underline{S}$ is in B then the m/d_i distinct cyclic permutation of $\underline{S}$ must also be in B. It is obvious that a state $\underline{S}$ can only be on a single closed cycle. Hence the set B may be partitioned into disjoint subsets, each containing m/d_i distinct states which are cyclic permutations of each other. The total number of such subsets is clearly equal to $(d_i/m)N^*(m,w,d_i)$. Q.E.D.

Summing equation (6) over the set D_w will yield the total number of distinct potential closed cycles of a given Hamming weight. Hence the following theorem.

Theorem 2

There are exactly $Z_w(m,w)$ distinct potential closed cycles of Hamming weight w on an MFSR of order M where

$$Z_w(m,w) = \sum_{d_i \varepsilon D_w} Z_d(m,w,d_i) = \sum_{d_i \varepsilon D_w} (d_i/m)N^*(m,w,d_i) \tag{7}$$

Finally summing equation (7) over the Hamming weight w, $o<w<m$, yields the total number of potential closed cycles. Thus:

Theorem 3

There are exactly Z(m) distinct potential closed cycles on an MFSR of order m where

$$Z(m) = \sum_{w=1}^{m-1} Z_w(m,w) = \sum_{w=1}^{m-1} \sum_{d_i \varepsilon D_w} (d_i/m)\, N^*(m,w,d_i) \qquad (8)$$

Corollary 1

If the order m is a prime number p, then

$$Z(m) = Z(p) = \frac{2^p-2}{p} \qquad (9)$$

Proof:

$$2^p = \binom{p}{o} + \sum_{w=1}^{p-1} \binom{p}{w} + \binom{p}{p}$$

$$= \sum_{w=1}^{p-1} \binom{p}{w} + 2$$

Therefore

$$\frac{2^p-2}{p} = \sum_{w=1}^{p-1} \frac{\binom{p}{w}}{p}$$

Since p is a prime number, $D_w = \{1\}$ for all w and

$$Z_w(p,w) = \frac{\binom{p}{w}}{p}$$

$$\sum_{w=1}^{p-1} Z_w(p,w) = \sum_{w=1}^{p-1} \frac{\binom{p}{w}}{p} = \frac{2^p-2}{p}$$

Q.E.D.

The total number of possible cycles on an MFSR of order m is equal to the number of possible closed cycles plus the number of possible equilibrium states. Denoting the number of possible cycles by $Z^*(m)$ one obtains

$$Z^*(m) = Z(m) + 2. \qquad (10)$$

Table 1[2] lists the number of potential closed cycles, T>1 for values of m from 1 to 24.

TABLE I

m	Z(m)	m	Z(m)	m	Z(m)	m	Z(m)
1	0	7	18	13	630	19	27594
2	1	8	34	14	1180	20	52486
3	2	9	58	15	2190	21	99878
4	4	10	106	16	4114	22	190744
5	6	11	186	17	7710	23	364722
6	12	12	350	18	14600	24	699250

Table II gives the potential closed cycle structure for an MFSR of order 6.

---

[2] All necessary computations were carried using the UNIVAC 1107 computer at the University of Notre Dame Computer Center.

TABLE II

w	$Z_d(6,w,1)$, T=6	$Z_d(6,w,2)$, T=3	$Z_d(6,w,3)$, T=2	$Z_\omega(6,w)$
1	1			1
2	2	1		3
3	3		1	4
4	2	1		3
5	1			1

$Z(6) = 12$ $\qquad$ $Z^*(6) = 14$

ENUMERATION OF STABLE MFSR'S

Definition 3

An MFSR is stable [4] if and only if for every state $\underline{S}$ there exists an integer K such that $\theta^K\underline{S}=\underline{0}$.

Definition 4

A state $\underline{S}$ on an MFSR of order m such that $S_0=1$ and $F(\underline{S})=0$ will be called a Hamming drop state since $W(\theta\underline{S})<W(\underline{S})$.

It is obvious that a necessary and sufficient condition for an MFSR to be stable is for $F(\underline{1})=0$ and for every potential closed cycle to contain at least one Hamming drop state.

The number of possible Hamming drop states on a potential closed cycle of period T and degeneracy d is equal to $\frac{w}{d}$ where w is the Hamming weight of the states which lie on the potential closed cycle. The only way a closed cycle can exist is for the potential cycle to contain no Hamming drop states. Hence, for each potential closed cycle there will be $2^{\frac{w}{d}}-1$ different ways of having at least one Hamming drop state.

The total number of stable MFSR's will be equal to the product of the number of different ways of having at least one Hamming drop state on each potential closed cycle, since there is only one way for $F(\underline{1})=0$. Thus:

Theorem 4

The number of stable MFSR's of order m denoted by K(m) is exactly

$$K(m) = \prod_{w=1}^{m-1} \prod_{d_i \epsilon D_w} \left(2^{\frac{w}{d_i}}-1\right)^{Z_d(m,w,d_i)} \tag{11}$$

where $D_w=\{d_i \mid d_i$ divides both m and w, $d_{i+1}<d_i\}$ and $Z_d(m,w,d_i)$ is given by equation 6.

Corollary 3

If the number of stages m is a prime number p then K(p) is given by

$$K(p) = \prod_{w=1}^{p-1} (2^w - 1)^{\binom{p}{w}/p} \tag{12}$$

Table III lists the number of stable MFSR's as well as the per cent stable for several values of m.

TABLE III

m	Z(m)	Z*(m)	K(m)	% Stable
1	0	2	1	100.
2	1	3	1	25.
3	2	4	3	18.75
4	4	6	21	8.2
5	6	8	6615	10.
6	12	14	64,595,475	1.5
7	18	20	$6.468\text{x}10^{17}$	3.5
8	34	36	$1.046\text{x}10^{36}$	.31
9	58	60	$2.247\text{x}10^{74}$	.19
10	106	108	$1.750\text{x}10^{150}$	.013
11	186	188	$2.082\text{x}10^{304}$	.0116
12	350	352	$2.391\text{x}10^{609}$	$7.4\text{x}10^{-6}$

A METHOD FOR DETERMINING STABILITY

Turning now to the investigation of the stability of an arbitrary MFSR, a method is presented by which its stability can be determined.

It has been stated that an MFSR is stable if and only if $F(\underline{1})=0$ and every potential closed cycle contains at least one Hamming drop state. Since the period T of any closed cycle is a divisor of the register length m, there can be no more than m states on any potential closed cycle and the following lemma is evident:

Lemma 3

An MFSR is stable if and only if every state except the $\underline{0}$ state is either a Hamming drop state or the i-th predecessor of a Hamming drop state for $o<i<m$.

The conditions for a Hamming drop state, $s_0=1$ and $F(\underline{S})=s_0\Phi(\hat{\underline{S}})=0$, $\hat{\underline{S}}=(s_1,s_2,\ldots,s_{m-1})$, imply $\Phi(\hat{\underline{S}})=0$. Denoting the function $S_0\Phi'(\hat{\underline{S}})$ by $\psi(\underline{S})$, the set A_0 of Hamming drop states consists of those states which satisfy the equation

$$\psi(\theta^o\underline{S}) \equiv \psi(\underline{S}) \equiv S_0\Phi'(\hat{\underline{S}}) = 1 \tag{13}$$

where the prime denotes binary complementation. The set A_1 of the first predecessors consists of those states whose successors after one shift satisfy equation 13, i.e., $\theta A_1=A_0$. Thus the states in A_1 satisfy

$$\psi(\theta^1\underline{S}) \equiv S_1\Phi'(\theta^1\hat{\underline{S}}) = 1 \tag{14}$$

In general, the set A_i of i-th predecessors consists of those states satisfying

$$\psi(\theta^i \underline{S}) \equiv S_i \Phi'(\theta^i \hat{\underline{S}}) = 1 \tag{15}$$

i.e., $\theta^i A_i = A_0$.

In effect, Lemma 3 states that an MFSR is stable if and only if every state except state $\underline{0}$ is an element of some A_i, $o \leq i < m$.

Theorem 5

An m-stage MFSR is stable if and only if the stability function $\Gamma(\underline{S})$,

$$\Gamma(S) = \bigvee_{i=o}^{m-1} \Psi(\theta^i \underline{S}) \vee \prod_{i=o}^{m-1} S_i \equiv 1 \tag{16}$$

Proof

If the MFSR is stable then every state $\underline{S}$ except the state $\underline{0}$ is an element of some A_i, $o \leq i < m$, and hence satisfies

$$\psi(\theta^i \underline{S}) = 1$$

Therefore all such states also satisfy

$$\bigvee_{i=o}^{m-1} \psi(\theta^i \underline{S}) = 1$$

If $\underline{S}$ is the $\underline{0}$ state then it satisfies

$$\prod_{i=0}^{m-1} S_i' = 1$$

Conversely, if every state satisfies equation (16), it is either the $\underline{0}$ state or an element of some A_i, $o \leq i < m$ and hence the shift register is stable Q.E.D.

Since the feedback function often involves modulo-2 addition rather than logical addition, an alternate form of the stability function using the modulo-2 (ring) sum is desired. It can be obtained as follows:

Since $\bigvee_{i=0}^{m-1} \Psi(\theta^i \underline{S})$ and $\prod_{i=o}^{m-1} S_i'$ can never be equal to 1 simultaneously, equation (16) can be rewritten as

$$\Gamma(\underline{S}) = \bigvee_{i=o}^{m-1} \Psi(\theta^i \underline{S}) \oplus \prod_{i=o}^{m-1} S_i' \equiv 1 \tag{17}$$

where $\oplus$ denotes modulo-2 addition.

By complementing the expression $\bigvee_{i=o}^{m-1} \Psi(\theta^i \underline{S})$ the entire left hand side of equation (17) is complemented. Thus, Theorem 5a. An m-stage MFSR is stable if and only if the associated stabilty function $\Gamma^*(\underline{S})$

$$\Gamma^*(S) = \prod_{i=0}^{m-1} \Psi'(\theta^i \underline{S}) \oplus \prod_{i=0}^{m-1} S_i' \equiv 0 \qquad (18)$$

Q.E.D.

Example. Consider the MFSR of order 5 defined by

$$F(\underline{S}) = s_0(1 \oplus s_1 \oplus s_2)(s_3 \oplus s_4) = s_0(s_1s_2 + s_1's_2')(s_3s_4' + s_3's_4)$$

The ψ functions for this MFSR are

$$\psi(\underline{S}) = s_0((1 \oplus s_1 \oplus s_2)(s_3 \oplus s_4) \oplus 1) = s_0(s_1s_2' + s_1's_2 + s_3s_4 + s_3's_4')$$

$$\psi(\ \underline{S}) = s_1(s_4(1 \oplus s_2s_3 \oplus s_0s_2s_3) \oplus s_0s_2s_3 \oplus 1)$$

$$= s_1(s_3's_4' + s_4's_0' + s_2's_3 + s_2s_3')$$

$$\psi(\theta^2\underline{S}) = s_2(1 \oplus s_1s_3s_4) = s_2(s_1' + s_3' + s_4')$$

$$\psi(\theta^3\underline{S}) = s_3$$

$$\psi(\theta^4\underline{S}) = s_4$$

Applying theorem 5 one obtains

$$\bigvee_{i=0}^{4} \psi(\theta^i\underline{S}) \vee \prod_{i=0}^{4} S_i' = s_0s_1s_2' + s_0s_1's_2 + s_0s_3s_4 + s_0s_3's_4' + s_1s_3's_4'$$

$$+ s_1s_4's_0' + s_1s_2's_3 + s_1s_2s_3' + s_2s_1'$$

$$+ s_2s_3' + s_2s_4' + s_3 + s_4 + s_0's_1's_2's_3's_4'$$

$$= s_0 + s_1 + s_2 + s_3 + s_4 + s_0's_1's_2's_3's_4'$$

$$= 1$$

Thus for m=5 the MFSR's defined by $F(\underline{S}) = s_0(1 \oplus s_1 \oplus s_2)(s_3 \oplus s_4)$ is stable. This MFSR is one for which the stability criterion of Massey and Liu[1] could not be applied.

REFERENCES

1. J. L. Massey & R. W. Liu, "Monotone Feedback Shift-Registers," Proc. 2nd Annual Allerton Conference on Circuit & System Theory, Univ. of Illinois, pp. 860-874, Sept., 1964.

2. S. W. Golomb & L. R. Welch, & R. M. Goldstein, "Cycles from Nonlinear Shift Registers," JPL Prog. Rept. No. 20-389, Calif. Inst. of Technology, Aug. 31, 1959.

3. B. Elspas, "The Theory of Autonomous Linear Sequential Networks," IRE Trans. on Circuit Theory, Vol. CT-6, March 11, 1959, pp. 45-60.

4. J. L. Massey & R. W. Liu, "Application of Liapunov's Direct Method to The Error-Propagation Effect in Convolutional Codes," IEEE Trans. on Info. Theory (Correspondence), Vol. IT-10, July, 1964, pp. 248-250.

SEQUENTIAL CODES AND DISCRETE MEMORYLESS CHANNELS WITH AN ERASURE PROPERTY*

TERRY J. WAGNER
Department of Electrical Engineering and Laboratories for
Electronics and Related Science Research, The University of Texas
Austin, Texas

Summary. The following question is asked. Are there codes for discrete memoryless channels, other than block codes, for which a Shannon-type coding theorem is true. In the following a class of codes, called sequential codes, is described and it is shown that channel capacity for these codes is nonzero for discrete memoryless channels with an erasure property.

Notation. For any set A, A^n denotes the set of all n-tuples from A (n is a positive integer) and A^I denotes the set of all sequences from A. If A is finite, $|A|$ denotes the number of elements in A.

A discrete memoryless channel (DMC), its n-extension, and its capacity will be denoted by (A,B,p), (A^n,B^n,q), and Q, respectively. The product of the DMC's (A,B,p) and (A',B',p') is the DMC $(A\times A',B\times B',q)$ where $q(y,y'/x,x') = p(y/x)p'(y'/x')$ for all $x \in A$, $x' \in A'$, $y \in B$, $y' \in B'$. A source for a DMC will be denoted by $\{Z_i\} = \{Z_i, -\infty < i < \infty\}$, a sequence of finite-valued random variables which are independent and identically distributed with a uniform distribution. It may be helpful to think of the subscript i in $\{Z_i\}$ as corresponding to time in seconds.

All random variables, unless otherwise noted, will be finite-valued. The ranges of X, Y, and Z will, with or without affixes, be denoted by A, B, and C, respectively. Thus X(Y) will, with a subscript to distinguish order, denote an input (output) random variable to the DMC (A,B,p). Finally, all logarithms are to the base 2.

Sequential Coding. A sequential encoder (SE) for a source $\{Z_i\}$ and a DMC (A,B,p) is a function $f^{(m)}: C^m \to A$ where (i) m is a positive integer called the duration of memory of the encoder and (ii) the channel inputs are related to source outputs by

---

*This work was partially supported by contract Nonr-222(53) with the Office of Naval Research, by the NSF Grant GP-3012, and the Joint Services Electronics Program under Grant AF-AFOSR 766-65.

$$X_j = f^{(m)}(Z_{j-m+1},\ldots Z_j),\ -\infty < j < \infty \qquad (1)$$

$f^{(m)}$ is frequently called a convolutional encoder [1,2] when it is linear in the sense of some finite field appropriate to C and A.

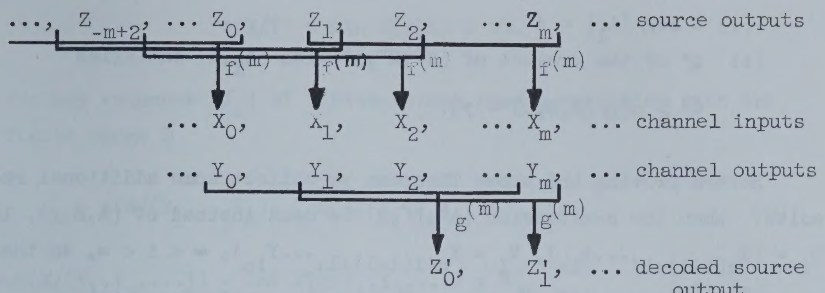
For a sequential encoder $f^{(m)}$, the decision for source output Z_1, here denoted by Z_1', is usually made [1,2] from the channel outputs $Y_1,\ldots Y_m$ assuming that $Z'_{-m+2},\ldots Z_0'$ are correct. Then Z_2' is decided on from $Y_2,\ldots Y_{m+1}$ assuming $Z'_{-m+3},\ldots Z_1'$ are correct and so on for $Z_3'\ldots$. The whole coding procedure is made into a block code by periodically putting in m-1 source symbols known to the decoder. This process, referred to as clearing or resynchronization, guarantees that the decoder can periodically start decoding a block of source random variables knowing exactly the previous m-1 source symbols. At present the main motivation for using sequential encoding is that there are a number of decoding methods (all of the methods use resynchronization) which have attractive implementation features [1,2,3,4,5]. Here, however, the interest will be in using a decoder which does not use resynchronization.

For a given $f^{(m)}$ suppose in decoding Z_1 we only use $Y_1,\ldots Y_m$, i.e., suppose we use a function $g^{(m)}: B^m \to C$ such that the channel input and output random variables are related to the source and decoded source random variables by

$$X_i = f^{(m)}(Z_{i-m+1},\ldots Z_i),$$

$$Z_i' = g^{(m)}(Y_i,\ldots Y_{i+m-1}),\ -\infty < i < \infty. \qquad (2)$$

The encoder and decoder are illustrated in Figure 1. $Y_1,\ldots Y_m$ are used to decode Z_1 since, in the encoder, $X_1,\ldots X_m$ are the only channel inputs which depend on Z_1 and, intuitively, $Y_1,\ldots Y_m$ should be the most important channel outputs for decoding Z_1. Of course using other channel outputs will improve the probability of decoding Z_1 correctly but, for the present, only $Y_1,\ldots Y_m$ will be used. We call $g^{(m)}$ a sequential decoder (SD) of memory duration m. The average numbers of bits sent per channel input is the rate of the code $(f^{(m)},g^{(m)})$ and equals $\log |C|$. Because of the stationarity involved with the above codes it is clear that $P[Z_i' \neq Z_i / f^{(m)},g^{(m)}]$, the probability that $Z_i' \neq Z_i$ when a SC $(f^{(m)},g^{(m)})$ is used, is the same for all i. Thus it suffices to restrict attention to Z_1 when discussing the probability of error for Z_i.



$\dots, Z_{-m+2}, \dots Z_0, Z_1, Z_2, \dots Z_m,$ ... source outputs
$f^{(m)}$
$f^{(m)}$
$f^{(m)}$
$f^{(m)}$
$\dots X_0, X_1, X_2, \dots X_m,$ ... channel inputs
$\dots Y_0, Y_1, Y_2, \dots Y_m,$ ... channel outputs
$g^{(m)}$
$g^{(m)}$
$Z_0', Z_1',$ ... decoded source output

Figure 1. Coding Relations for the SC code $(f^{(m)},g^{(m)})$ of (2).

When a SC $(f^{(m)},g^{(m)})$ is used for $\{Z_i\}$ and (A^n,B^n,q) the rate R of the code *referred to channel (A,B,p)* is $\log |C|/n$, the average number of bits sent per input of (A,B,p). Motivated by Shannon's coding theorem for block codes with DMC's [6] we ask when does there exist a sequence of SC's $\{(f^{(m)},g^{(m)}), m=1,2,\dots\}$ for a source $\{Z_i\}$ and (A^n,B^n,q) such that $P[Z_i' \neq Z_1/f^{(m)},g^{(m)}] \overset{m}{\to} 0$. Corresponding to the channel capacity Q of the DMC (A,B,p) we define $Q^* = Q^*(A,B,p)$ as the supremum of rates $\log |C|/n$ such that there exists a sequence of SC's $\{(f^{(m)},g^{(m)})\}$ for $\{Z_i\}$ and (A^n,B^n,q) with $P[Z_1' \neq Z_1/f^{(m)},g^{(m)}] \overset{m}{\to} 0$. Q^* is called the sequential coding capacity.

Unfortunately, what Q^* is for a DMC (A,B,p) is an open question. Below a class of channels is described for which $Q^* > 0$. We mention at this point that it is straightforward to show that $n\,Q^*(A,B,p) = Q^*(A^n, B^n,q)$. Furthermore, Shannon's result [7, theorem 1] implies that $Q^* \leq Q$. (A more explicit proof may be found in [8]).

For any finite set A and any δ with $0 \leq \delta \leq 1$ the *erasure channel* (A,δ) is the DMC $(A, A \cup \{s\}, p)$ where s is the output which occurs when the input is erased and $p(x/x) = \delta$, $p(s/x) = 1 - \delta$ for all $x \in A$. For an erasure channel (A,δ), $Q = \delta \log |A|$.

The main results can now be stated.

Theorem 1. Any DMC (A_1,B_1,p_1) which is the product of an arbitrary DMC (A,B,p) and an erasure channel (A_2,δ) has $Q^*(A,B,p) > 0$.

Theorem 2. For any DMC (A,B,p) and any $\epsilon > 0$, $\gamma > 0$ there is an integer n and an erasure channel (A_1,δ) with the following properties:

(i) $\delta \log |A_1| < \epsilon$

(ii) Q^* of the product of (A^n, B^n, q) with (A_1, δ) satisfies

$$Q^* \geq n[Q(A,B,p) - \gamma].$$

Before proving the above Theorems we collect some additional results. When the n-extension (A^n, B^n, q) is used instead of (A,B,p), let $U_i = (X_{n(i-1)+1}, \ldots X_{in})$, $V_i = Y_{n(i(i-1)+1,} \ldots Y_{in})$, $\infty < i < \infty$, so that a SC $(f^{(m)}, g^{(m)})$ for (A^n, B^n, q) and $\{Z_i\}$ has the following notation:

$$U_i = f^{(m)}(Z_{i-m+1}, \ldots Z_i)$$
$$Z_i' = g^{(m)}(V_i, \ldots V_{i+m-1}) \tag{3}$$

Denote the minimum probability of error over all ways of deciding Z_1 from $(Z_a, \ldots Z_b;\ V_c, \ldots V_d)$ $(a \leq b,\ c \leq d)$ when a SE $f^{(m)}$ is used by $e(Z_1/f^{(m)};\ Z_a, \ldots Z_b;\ V_c, \ldots V_d)$. Also, let $e(Z_1/f^{(m)};\ V_c, \ldots V_d)$ be the minimum probability of error over all ways of deciding Z_1 from $(V_c, \ldots V_d)$. For example, for $c = 1$ and $d = m$, $e(Z_1/f^{(m)};\ V_1, \ldots V_m) = \min\limits_{g^{(m)}} P[Z_1' \neq Z_1/f^{(m)}, g^{(m)}]$.

Reiffen [2] has shown that for $R = \log |C|/n < Q$, there exists a sequence of sequential encoders $\{f_o^{(m)}\}$ such that for all m

$$e(Z_1/f_o^{(m)};\ Z_{-m+2}, \ldots Z_0;\ V_1, \ldots V_m) \leq L(m) 2^{-mnE(R)} \tag{4}$$

where $L(m)$ is a polynomial in m (whose coefficients depend on n) and $E(R) > 0$ for all $R < Q$. Before making use of (4) three lemmas are given below, the first two of which are routine.

Lemma 1. For a SE $f^{(m)}$ and a positive integer r $(m \geq 2)$

$$e(Z_1/f^{(m)};\ Z_{-m-r+3}, \ldots Z_{-r+1};\ V_{-r+2}, \ldots V_m) \leq$$
$$r e(Z_1/f^{(m)};\ Z_{-m+2}, \ldots Z_0;\ V_1, \ldots V_m). \tag{5}$$

Lemma 2. For any finite-valued random variables X and Y let

$$e(X/Y) = \min_{\{g:\ B\to A\}} P[g(Y) \neq X].$$

Then for any sequence $\{Y_i\}$ of finite-valued random variables with the same finite range B

$$e(X/(Y_1,\ldots Y_n)) \overset{n}{\downarrow} e(X/(Y_1,Y_2,\ldots)), \qquad (6)$$

where $e(X/(Y_1,Y_2,\ldots)) = \inf\limits_{\{g:\ g^{-1}(x)\, \in\, \sigma\,(B^I)\text{ for all } x\}} P[g(Y_1,Y_2,\ldots) \neq X]$,

and $\sigma(B^I)$ is the product σ-field of sets of B^I determined by cylinder sets (see [9] for definitions).

Let $e(Z_1/f^{(m)};\ V_1,\ldots V_m) =$

$$\inf_{\{g:\ g^{-1}(x)\, \in\, \sigma(B^I)\text{ for all } x\}}\{P[g(\ldots V_1,\ldots V_m) \neq Z_1] \text{ when } f^{(m)} \text{ is the encoder}\}.$$

Lemma 3. Suppose there exists a sequence of SE's $\{f_o^{(m)}\}$ for $\{Z_i\}$ and (A^n,B^n,q) such that $e(Z_1/f_o^{(m)};\ldots V_1,\ldots V_m) \xrightarrow{m} 0$. Then there exists a sequence $\{(f^{(m)},g^{(m)})\}$ such that $P[Z_i' \neq Z_i/f^{(m)},g^{(m)}] \xrightarrow{m} 0$.

Proof. Let $a_{m,r} = \min\limits_{\{\text{all SE's } f'^{(m)} \text{ of memory } m\}} \{e(Z_1/f'^{(m)};\ V_{m-r+1},\ldots V_m)\}$, $m,r = 1,2\ldots$

Then (i) $a_{m,r} \geq a_{m,r+1}$ and, from Lemma 2,

$$\lim_{r\to\infty} a_{m,r} \leq e(Z_1/f_o^{(m)};\ldots V_1,\ldots V_m) \text{ so that}$$

$$\lim_{m\to\infty}\lim_{r\to\infty} a_{m,r} = 0.$$

(ii) $a_{m,r} \geq a_{m+1,r}$.

To prove (ii) notice that for $f_r^{(m)}$, the SE which yields $a_{m,r} = e(Z_1/f_r^{(m)};\ V_{m-r+1},\ldots V_m)$, the SE $h^{(m+1)}$ defined by

$$V_j = h^{(m+1)}(Z_{j-m},\ldots Z_j) = f_r^{(m)}(Z_{j-m},\ldots Z_{j-1})$$

merely advances the V's one coordinate so that the probability distribution of $(Z_1;\ V_{m-r+1},\ldots V_m)$ with $f_r^{(m)}$ is the same as $(Z_1;\ V_{m-r+2},\ldots V_{m+1})$

with $h^{(m+1)}$. Thus $a_{m+1,r} \le e(Z_1/h^{(m+1)};\ V_{m-r+2},\dots V_{m+1}) = a_{m,r}$.

From (i) and (ii) it follows that $a_{m,m} \overset{m}{\to} 0$ so that the sequence $\{f_m^{(m)}\}$ of SE's has $e(Z_1/f_m^{(m)};\ V_1,\dots V_m) \overset{m}{\to} 0$. Since there is always a SD $g^{(m)}$ which achieves $e(Z_1/f_m^{(m)};\ V_1,\dots V_m)$ the proof is completed.

Proof of Theorem 1. From (4) when $R = \log |C| < Q$ there exists a sequence $\{f_o^{(m)}\}$ of SE's for $\{Z_i\}$ and (A^n,B^n,q) where, for all m,

$$e(Z_1/f_o^{(m)};\ Z_{-m+2},\dots Z_0;\ V_1,\dots V_m) \le L(m)2^{-mnE(R)}.$$

Choose n so that there is an integer k, $k \le n$, with $|C| \le |A_2|^k$. Thus there is mapping $f\colon C \to A_2^n$ which is 1-1 on the first k coordinates of A_2^n. Furthermore,

$$U_i' = (f(Z_i), f_o^{(m)}(Z_{i-m+1},\dots Z_i)),\ -\infty < i < \infty \tag{7}$$

is a SE $f^{(m)}$ of memory duration m for $\{Z_i\}$ and (A_1^n,B_1^n,q_1), the n-extension of (A_1,B_1,p_1). In (7) U_i' denotes the ith input random variable to (A_1,B_1,q_1) where, for convenience, we have identified (A_1,B_1,q_1) with the product of (A^n,B^n,q) and the n-extension of (A_2,δ).

Let V_i' denote the ith output random variable for (A_1^n,B_1^n,q_1). We now show that if n and k satisfy $[nE(R)-k\log(1/\delta)] > 0$, then $e(Z_1/f^{(m)};\ \dots V_1',\dots V_m') \overset{m}{\to} 0$ which, by Lemma 3, implies that $Q^*(A_1,B_1,p_1) \ge \log |C|/n > 0$.

Decoding Z_1 from the sequence $(\dots V_1',\dots V_m')$ is accomplished as follows. Starting with V_0' and working backward, m-1 consecutive outputs, say $V'_{-r+1},\dots V'_{-r+m-1}$ $(r \ge m-1)$, are sought where the first k erasure channel outputs of each V_i', $i = -r+1,\dots -r+m-1$, are unerased. By construction of $f^{(m)}$, $Z_{-r+1},\dots Z_{-r+m-1}$ are then known (f was 1-1 on the first k erasure channel inputs). From Lemma 1 Z_1 can then be decoded from $V'_{-r+1},\dots V_m'$ with a probability of errors $\le (r-m+2)L(m)2^{-mnE(R)}$. Let $p_{m,r}$ be the probability of finding m-1 consecutive outputs as described above for the first time at coordinate $-r+1$. Then

$$e(Z_1/f^{(m)};\ \dots V_1',\dots V_m') \le \sum_{r=1}^{\infty} (r-m+2)p_{m,r} L\,(m)2^{-mnE(R)}$$

$$\leq \sum_{r=1}^{\infty} rp_{m,r} L(m) 2^{-mnE(R)} \quad (\text{Assume } m \geq 2)$$

$$= \frac{1 - (\delta^k)^{m-1}}{(1 - \delta^k)(\delta^k)^{m-1}} L(m) 2^{-mnE(R)}$$

$$\leq L(m) \frac{\delta^k}{1 - \delta^k} 2^{-m[nE(R) - k\log(1/\delta)]}$$

$$\overset{m}{\rightarrow} 0 \text{ if } [nE(R) - k\log(1/\delta)] > 0.$$

The above equality follows from Feller [10], page 300, with Feller's p, q, r replaced by δ^k, $1-\delta^k$, and m-1, respectively. This completes the proof.

Proof of Theorem 2. Consider the following: (1) a source $\{Z_i\}$ where C is the set of all k-tuples of 0's and 1's, and (2) an erasure channel (A_1,δ) where $A_1 = C$ and $\delta = \mu^k$. As in the proof of Theorem 1, if $k/n < Q$ there exists a sequence of SE's $\{f_o^{(m)}\}$ for $\{Z_i\}$ and (A^n,B^n,q) such that

$$e(Z_1/f_o^{(m)};\ Z_{-m+2},\ldots Z_o;\ V_1,\ldots V_m) \leq L(m)2^{-mnE(R)}$$

for all m. Furthermore, $U_i' = (Z_i, f_o^{(m)}(Z_{i-m+1},\ldots Z_i))$, $-\infty < i < \infty$, is a sequential encoder $f_o^{(m)}$ of memory duration m for $\{Z_i\}$ and the product of (A^n,B^n,q) with (A_1,δ). By the same argument used in the proof of Theorem 1 we find that Q* for the product is $\geq k = Rn$ if $[nE(R) - k\log(1/\mu)] > 0$. It is clear that if $[nE(R) - k\log(1/\mu)] > 0$ for some $n = n_o$, $k = k_o$ and $R = k_o/n_o$ then it is greater than zero for all k, n with $R = k/n$.

Select $R = k_o/n_o$ such that $Q > R \geq Q - \gamma$. Next select $\mu > 0$ such that $n_oE(R) - k_o\log(1/\mu) > 0$. Since the channel capacity of (A_1,δ) is $k\mu^k = \delta \log |A_1|$, it follows that we can find an erasure channel (A_1,δ) with capacity $< \epsilon$ by choosing k and n sufficiently large with $k/n = k_o/n_o = R$. This completes the proof.

Aside from calculating Q* for a given DMC it would be nice to know if Q* > 0 for a channel which does not have the erasure properties used above, for example, is Q* > 0 for the Binary Symmetric Channel.

Another question arises from the fact that Lemma 3 does not provide any information about how fast $P[Z_1' \neq Z_1/f^{(m)}, g^{(m)}]$ tends to 0 with m. For example, when

$$e(Z_1/f^{(m)}; \ldots V_1', \ldots V_m') \leq L(m) \frac{\delta^k}{1-\delta^k} 2^{-m[nE-k\log(1/\delta)]},$$

and hence tends to 0 exponentially with m, does $e(Z_1/f^{(m)}; V_1', \ldots V_m')$ tend to 0 exponentially with m? More generally one would like to know how much smaller $e(Z_1/f^{(m)}; V_{-r'+2}, \ldots V_{m+r''-1})$ is than $e(Z_1/f^{(m)}; V_1, \ldots V_m)$ for r', r" = 1,2,... .

Acknowledgment. The writer wishes to express his appreciation to Professor A. J. Thomasian for his guidance during the course of the work leading to this report.

REFERENCES

1. J. M. Wozencraft and B. Reiffen. (1961). Sequential Decoding. M.I.T. Press and John Wiley and Sons, Inc., New York.
2. B. Reiffen. (1960). Sequential Encoding and Decoding for the Discrete Memoryless Channel. M.I.T., Cambridge, Mass., Res. Lab. or Electronics, Tech. Rept. No. 374.
3. R. Fano. (1963). A Heuristic Discussion of Probabilistic Decoding. IEEE Trans. on Information Theory, vol. IT-9, pp. 64-73, April.
4. J. Ziv. (1963). Successive Decoding Scheme for Memoryless Channels. IEEE Trans. on Information Theory, vol. IT-9, pp. 97-104, April.
5. J. Ziv. (1962). Coding and Decoding for Time-Discrete, Amplitude Continuous Memoryless Channels. M.I.T., Cambridge, Mass., Res. Lab. of Electronics, Tech. Rept. No. 399.
6. C. E. Shannon. (1957). Certain Results in Coding Theory for Noisy Channels. Information and Control, vol. 1, no. 1, pp. 6-25, September.
7. C. E. Shannon. (1960). Coding Theorems for a Discrete Source with a Fidelity Criterion. Information and Decision Processes. McGraw-Hill Book Co., New York.
8. T. Wagner. (1962). Generalized Coding and Uniform Finite-Memory Codes. University of California, Berkeley, Electronics Research Laboratory, Series 60, No. 492, December.
9. M. Loève. (1960). Probability Theory. D. Van Nostrand Co., Princeton, New Jersey.
10. W. Feller. (1957). An Introduction to Probability Theory and Its Applications. John Wiley and Sons, Inc., New York.

FUNCTIONAL TRANSFORMATIONS FOR TESTING THE REALIZABILITY OF SINGLE-CONTACT FUNCTIONS

S. PARK CHAN and EDWARD R. MEYER
University of Santa Clara, Santa Clara, California

ABSTRACT

One of the important problems in the synthesis of single-contact (SC) networks is how to characterize SC functions so that one can determine, by simple means, whether or not a given function is an SC function. Several papers[1-4] exist which are concerned with the synthesis of switching networks. Also, sufficient conditions have been given for the realization of SC networks by testing a given function to determine if it satisfies the sufficient conditions. The procedures of the testings, however, usually involve the realization of either an incidence matrix[3] or a circuit matrix[1], and such procedures are rather complicated and time consuming.

This paper deals with the derivation of a set of functional transformations. Any of these transformations, when applied to a given function will transform the function into a simpler function which corresponds to another SC network if the original one is an SC function. Thus, these functional transformations serve to simplify the task of the synthesis. On the other hand, if the original function is not an SC function to begin with, then the applications of the functional transformations may reveal the nonrealizability of the function and, consequently, save time and labor.

INTRODUCTION

In this work, some functional transformations are introduced. These transformations are used to reveal the topological properties of a single-contact (SC) network through the investigation of the corresponding SC function.

All of the lemmas, theorems and corollaries included in this paper will be stated without proof because of the limited space allowed. The interested reader, however, may find the proofs in references 5 and 6.

SOME DEFINITIONS AND THEOREMS

We shall include, in this section, some definitions and theorems which are useful in developing the set of transformations to be discussed in the sections to follow. For a more detailed discussion on these definitions and theorems, as well as other definitions used in this paper, the reader may refer to reference 5.

First, two kinds of edge transformations--edge-opening transformation and edge-shorting transformation--will be defined.

Definition 1. Edge-opening transformation. Let $F(x_i=0)$ be the function obtained from a function F by setting $x_i=0$; then $F(x_i=0)$ is the edge-opening transformation of F with respect to x_i. Alternatively, we say that $F(x_i=0)$ is obtained from F by "opening" x_i in F.

It is obvious that any edge-opening transformation of an SC function is again an SC function because if we let the given function F be the SC function of some networkN, then, be definition, the edge-opening transformation $F(x_i=0)$ of F with respect to edge x_i is simply the SC function of the new network N' resulting from N by opening edge x_i.

Now, suppose edge x_i is shorted in N, then the endpoints of x_i will coincide and all the paths between the terminals of N containing x_i will no longer contain x_i. Each of these paths will become either a path between the terminals of the resultant network $N(x_i=1)$ or a "redundant path" in $N(x_i=1)$ in the sense that the edges contained in it form a path and a circuit in $N(x_i=1)$. In the latter case, the term corresponding to the path becomes a redundant term in the new function $F(x_i=1)$ which is realized by $N(x_i=1)$. Therefore, the new function $F(x_i=1)$, which is obtained by removing from F the variable x_i in each term containing it, and also removing the redundant terms, is the SC function of the resultant network $N(x_i=1)$ where the edge x_i has been shorted. Therefore, $F(x_i=1)$ can be obtained by setting $x_i=1$ and removing the redundant terms in F.

Definition 2. Edge-shorting transformation. Let $F(x_i=1)$ be the function which is obtained from a function F by first setting $x_i=1$ and then removing the redundant terms, if any, from the function. Then $F(x_i=1)$ is the edge-shorting transformation of F with respect to x_i. Alternatively, we say that $F(x_i=1)$ is obtained from F by "shorting" the variable x_i in F.

We now see that any edge-shorting transformation of an SC function is again an SC function.

Definition 3. Terminal transformation (with respect to some variable x_k). Let the function F of e variables be the SC function between vertices v_i and v_j of an SC network N of v vertices and e edges. Also, let $S_1=\{f_{1i}/i=1,2,\ldots\}$ be the set of all terms in F containing the variable x_k, and let $S_2=\{f_{2j}/j=1,2,\ldots\}$ be be the set of all terms in F not containing x_k. Then the function $F_v(x_k)$, which is obtained step-by-step by the application of the following operations, is called the terminal transformation of F with respect to x_k.

a) Obtain the function F' where $F'=x_k+\sum_{i,j} f_{1i}\oplus f_{2j}\oplus x_k$ when $S\neq\emptyset$,

$F'=x_k$ when $S=\emptyset$.

b) Remove all repeated and redundant terms in F' and the resultant function is $F_v(x_k)$.

Theorem 1. Let the function $F_v(x_k)$ of e' variables be the terminal transformation of a given SC function F of e variables of some SC network N. Then N is nonseparable if and only if e'=e.

As the consequence of Theorem 1, we have the following corollary.

Corollary 1. Let the function $F_v(x_k)$ of e' variables be the terminal transformation of a given SC function F of e variables of some SC network N. Then

a) $r' \leqq r$ always; and b) $r'=r$ if N is nonseparable,

where r and r' are the ranks of the path matrices corresponding to F and $F_v(x_i)$ respectively.

We shall now summarize the effects of the edge transformations and terminal transformations of an SC function of the original SC network. Let N be any given SC network whose function between vertices v_i and v_j is denoted by F which is of rank r. Also let F' stand for any of the functional transformations and let N' be an SC network realizing F'. Finally, let N have v vertices and e edges, and let N' have v' vertices and e' edges. Then,

a) If F' is an edge-opening transformation of F, then N' is a proper subnetwork of N. Also $v' \leqq v$, $e'<e$, and $r'<r$.

b) If N' is an edge-shorting transformation of F, then N' may or may not be (it generally is not) a subnetwork of N. Moreover,

c) If F' is a terminal transformation with respect to $x_i(v_i,v_j)$ and

i) If $e' \leqq e$, then N' is a subnetwork of N, $v' \leqq v$ and $r' \leqq r$;
ii) If $e'<e$, then N is separable, N' is a proper subnetwork of N containing x_i, $v'<v$, and $r' \leqq r$;
iii) If $e'=e$, then $N'=N$; that is, N' and N are identical.

SOME FUNCTIONAL TRANSFORMATIONS

In this section, a set of functional transformations will be introduced, Several theorems will also be stated as the consequence of these transformations. Their proofs can be found in reference 6.

Definition 4. Path-basic-cutset transformation. Let F_{AB} be a single contact switching function defined between the terminal vertices A and B of the graph G. Let F be expressed in natural form; i.e.,

$$F_{AB}=(a_1a_2\ldots a_m)+(b_1b_2\ldots b_n)+\ldots+(p_1p_2\ldots p_r) \text{ or } F_{AB}=\sum_{k=1}^{p} P_{AB_k}$$

where $\{P_{AB_1}\}=\{a_1,a_2,\ldots,a_m\}$; $\{P_{AB_2}\}=\{b_2,\ldots,b_n\}$; $\{P_{AB_p}\}=\{P_2,\ldots,P_p\}$ are edges of the path PRODUCTS $P_{AB_1},P_{AB_2},\ldots,P_{AB_p}$ respectively (p being the total number of paths between i and j), and where some edges may appear in more than one path (i.e., a_k and b_m may be the same edge). Let the function F^*_{AB} be formed from F_{AB} by interchanging the sums and products so that

$$F^*_{AB}=(a_1+a_2+\ldots+a_m)(b_1+b_2+\ldots+b_n)\ldots(p_1+p_2+\ldots+p_p)$$

which, in natural form, becomes

$$F^*_{AB}=(a_1b_1\ldots p_1)+(a_1b_1\ldots p_2)+\ldots+(a_mb_n\ldots p_p).$$

The function F^*_{AB} is called the path-basic-cutset transformation with respect to the SC function F_{AB}.

Theorem 2. The function F^*_{AB} described in Definition 4 is the sum of all basic cutsets of G; that is, eac h term in the sum represents a basic cutset of G between vertices A and B.

Corollary 2. If H is the union of basic cutsets of an SC network, then H^* is its switching function.

The following example will illustrate the ideas involved.

Example 1: Consider the SC function F

$$F=x_1x_4x_7+x_1x_4x_5x_6+x_1x_3x_5x_7+x_1x_3x_6+x_2x_3x_4x_7+x_2x_5x_7+x_2x_6$$

which corresponds to the SC network N as shown in Fig. 1

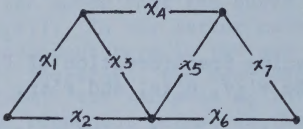


Figure 1

$$F^*=(x_1+x_4+x_7)(x_1+x_4+x_5+x_6)(x_1+x_3+x_5+x_7)(x_1+x_3+x_6)(x_2+x_3+x_4+x_7)$$
$$(x_2+x_5+x_7)(x_2+x_6)$$
$$=(x_1+x_4+x_5x_7+x_6x_7)(x_1+x_3+x_5x_6+x_6x_7)(x_2+x_7+x_3x_5+x_4x_5)(x_2+x_6)$$
$$= x_1x_2+x_2x_3x_4+x_2x_3x_5x_7+x_1x_3x_5x_6+x_4x_5x_6+x_6x_7.$$

By comparing each term of F^* to the basic cutsets of network N in Fig. 1, it is clear that F^* is the union of basic cutsets of network N.

At times it may be desirable to obtain the switching function of a dual of a given SC network. One reason is that the dual function may be an easier one to test but the main motivation is that, for every theorem about paths, there is a dual theorem about basic cutsets, and vice versa.

Definition 5. Duality transformation. Let G be a planar graph with a switching function $F(x_i)$ defined such that the vertices of edge x_i are the terminal vertices. Then $[x_i+\{x_i \oplus F^*(x_i)\}]|_{x_i=x'_i}$ is called the duality transformation of $F(x_i)$.

Theorem 3. The duality transformation defined above is the switching function defined between the vertices of x_i' of graph G' where G' is a dual of the graph G, and x_i' is the edge in G' corresponding to x_i in G.

Example 2: Consider the function F: $F=x_1x_4+x_1x_3x_5+x_2x_5+x_2x_3x_4$ which is realized by graph G in Fig. 2a.

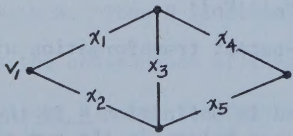


Figure 2a

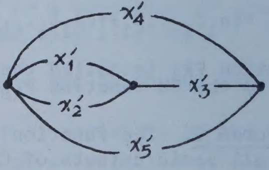


Figure 2b

$$F(x_1)=x_1+x_2x_3+x_2x_4x_5$$

$$F^*(x_1)=x_1(x_2+x_3)(x_2+x_4+x_5)=(x_1x_2+x_1x_3)(x_2+x_4+x_5)=x_1x_2+x_1x_3x_4+x_1x_3x_5$$
$$x_1\oplus F^*(x_1)=x_2+x_3x_4+x_3x_5; \{x_1+[x_1\oplus F^*(x_1)]\}\big|_{x_i=x'_i}=x'_1+x'_2+x'_3x'_4+x'_3x'_5$$

This is seen to be the switching function of G' across edge x'_i.

Next, a transformation will be presented such that if two switching functions are known--one between vertices A and C of a graph and the other between vertices B and C of the graph--then the transformation will yield the switching function between the vertices A and B.

Definition 6. Path transformation F_{AB}/C. A path transformation F_{AB}/C of the functions F_{AC} and F_{BC} with respect to vertex C where A, B, and C are vertices of the graph G is the function obtained from the following two operations:

a) Obtain the function F_{ABC} where $F_{ABC}=\sum_i \sum_j P_{AC_i}\oplus P_{BC_j}$; and where $P_{AC_i}\subseteq S_{AC}$, the set of all paths containing A and C as terminal vertices; and where $P_{BC_j}\subseteq S_{BC}$, the set of all paths containing B and C as terminal vertices.

b) Remove all repeated and redundant terms in F_{ABC}. The resultant function is F_{AB}/C.

Theorem 4. The path transformation F_{AB}/C is equal to F_{AB}, the switching function having A and B as terminal vertices.

The path transformation will now be used to develop another transformation. From the terminal cutset condition, it may be possible to obtain one or both of the terminal cutsets. Any two edges in a terminal cutset are incident to the same vertex (the terminal vertex), and the switching functions across these edges may be obtained through use of the terminal transformation. With these functions known, the switching function between the non-terminal vertices may be obtained from the path transformation.

Definition 7. Terminal-cutset-path transformation. Let F be an SC function realizable by graph G and let C be a terminal vertex (with respect to F) of G such that C is at least of degree 2. If x_i and x_j are edges incident to C, and A and B are the non-terminal vertices of x_i and x_j respectively, then $F(x_ix_j)$ is a terminal-cutset-path transformation and is formed by the following two operations:

a) Form $F'(x_ix_j)=\sum_k \sum_l F_k(x_i)\oplus F_l(x_j)$ where $F_k(x_i)$ is a term in $F(x_i)$, the terminal transformation with respect to x_i; and $F_l(x_j)$ is a term in $F(x_j)$, the terminal transformation with respect to x_i

b) Delete all redundant and repeated terms.

Theorem 5. $F(x_ix_j)$ is the sum of path products of G with A and B as terminal vertices.

Example 3: Consider the function F: F=ad+aef+bcde+bcf, which is realized by graph G in Fig. 3.

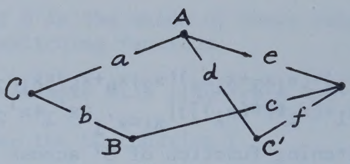


Figure 3

$F(a)=a+bce+cdf$ $F(b)=b+ace+acdf$

$F'(ab)=ab+ce+cdf+ce+ab+abdef+cdf+abdef+ab$ $F(ab)=ab+ce+cdf$

Definition 8. Terminal edge. A terminal edge is an edge belonging to a terminal cutset.

Theorem 6. Let F be an SC function realizable by G and let C and C' be the terminal vertices (with respect fo F) of G. Let x_i be a terminal edge with vertices C and A. The function $F_{AC'}/C$ contains a prime basic cutset containing the edge x_i.

Example 4: Consider the function F: $F=adg+adef+aceg+acf+bcdg+beg+bf$. {a,b} and {f,g} are the terminal cutsets of G if G exists. Let C be the terminal vertex incident to a and b, and let C' be the other terminal vertex and A be the other vertex of a.

$F_{AC}=F(a)=a+bc+bde+bdfg$

$F_{CC'}=F=adg+adef+aceg+acf+bcdg+beg+bf$

$F_{AC'}/C=dg+deg+ceg+cf+abeg+abf$

The prime basic cutsets are {a,b} and {f,g}. Therefore the necessary condition stated above is satisfied. F may be realized by the graph G as shown in Fig. 4.

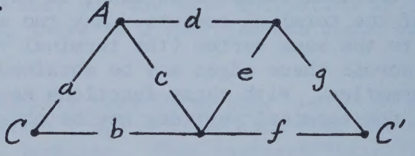


Figure 4

Example 5: Consider the function F: $F=abd+acr+bce+def$. {a,e}, {b,f} and {c,d} are prime basic cutsets of F if F is an SC function. Assume that {a,e} is a terminal cutset. Let C be the terminal vertex and A the other vertex of a.

$F_{AC}=F(a)=a+cde+bef$ $F_{CC'}=abd+acf+bce+def=F$

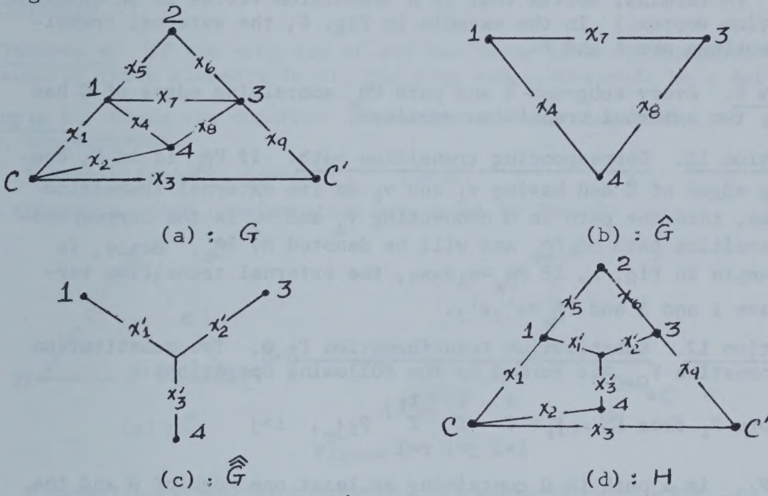
$F_{AC'}/C=bd+cf+abce+adef$

Since a is in at least one path with all other edges, a cannot be in a prime cutset of $F_{AC'}/C$. From these results, it is seen that {a,e} is not a terminal cutset. By similar development, it can be shown that {b,f} and {c,d} are not terminal cutsets and, therefore, F is not an SC function.

SUBGRAPH SUBSTITUTION

The objective in this section is to present a technique for obtaining the switching function of a graph G after it has been modified by replacing any subgraph $\hat{G}$ of G by a graph $\hat{\hat{G}}$ where $\hat{\hat{G}}$ is a star. Let an SC function F be realizable by graph G with C and C' taken as terminal vertices. Let $\hat{G}$ be a subgraph of G containing $\hat{v}$ vertices incident to both $\hat{G}$ and its complement such $\hat{v} < v$ where v is the number of vertices of the graph G. Vertices that are incident to both a subgraph and its complement shall be called the boundary vertices of the subgraph. Let graph G be modified to form graph H by removing $\hat{G}$ from G and replacing $\hat{G}$ with another graph $\hat{\hat{G}}$ with the property that $\hat{\hat{G}}$ has $\hat{v}+1$ vertices.

As an example, let G be the graph shown in Fig. 5, and let $\hat{G}$ and $\hat{\hat{G}}$ be defined respectively as the delta containing edges x_4, x_7 and x_8 in G, and $\hat{\hat{G}}$ be a star identified by edges x'_1, x'_2, x'_3 which do not belong to G.



(a) : G (b) : $\hat{G}$

(c) : $\hat{\hat{G}}$ (d) : H

Figure 5

In this example, the boundary vertices of $\hat{G}$ are 1, 3 and 4, which also happen to be the only vertices of $\hat{G}$. $\hat{\hat{G}}$ satisfies the condition that the number of vertices in $\hat{\hat{G}}$ be one greater than the number of boundary vertices in $\hat{G}$ and hence may be substituted for $\hat{G}$ forming the graph H as shown. In order to obtain all the paths in H, it will be necessary to define a set of vertices that is a subset of the boundary vertices.

Definition 9. Transition vertex. Let $P_{\hat{G}_k}$ be a path in graph G connecting the terminal vertices of G and containing edges of subgraph $\hat{G}$. A transition vertex of $\hat{G}$ and $P_{\hat{G}_k}$ is a vertex v in path $P_{\hat{G}_k}$ satisfying one of the following properties: 1) v is incident to one edge of $P_{\hat{G}_k}$ contained in $\hat{G}$ and one edge of $P_{\hat{G}_k}$ contained in the complement of $\hat{G}$; 2) if a terminal vertex is contained in $\hat{G}$, then it is a transition vertex.

The examples in Fig. 6 will clarify the meaning of the definition.

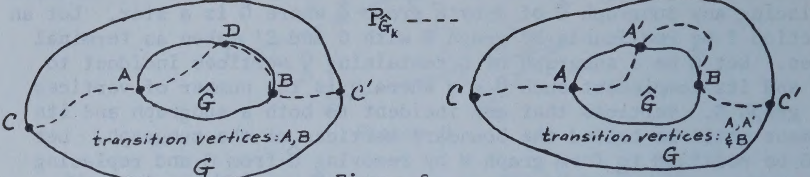


Figure 6

Definition 10. External transition vertex. A transition vertex is an external transition vertex of G and $P_{\hat{G}_k}$ if a path can be drawn in $P_{\hat{G}_k}$ to a terminal vertex of G that contains no other transition vertices. (A terminal vertex that is a transition vertex is an external transition vertex.) In the example in Fig. 6, the external transition vertices are A and B.

Theorem 6. Every subgraph $\hat{G}$ and path $P_{\hat{G}_k}$ containing edges of G has exactly two external transition vertices.

Definition 11. Corresponding transition path. If $P_{\hat{G}_k}$ is in G, containing edges of $\hat{G}$ and having v_1 and v_2 as its external transition vertices, then the path in G connecting v_1 and v_2 is the corresponding transition path of $P_{\hat{G}_k}$ and will be denoted by $P_{\hat{\hat{G}}_k}$. Hence, in the example in Fig. 5, if $P_{\hat{G}_k}=x_1x_7x_9$, the external transition vertices are 1 and 3 and $P_{\hat{\hat{G}}_k}=x'_1x'_2$.

Definition 12. Substitution transformation $F_{\hat{G}\to\hat{\hat{G}}}$. The substitution transformation $F_{G\to G}$ is formed by the following operations:

a) Form F_1 from F. $F_1 = \sum_{i=1}^{v}\sum_{j=1}^{v}\sum_{r=1}^{t_{ij}} P_{ij_r}$, $i>j$

where P_{ij_r} is a path in G containing at least one edge of G and the boundary vertices i and j of G, and where t_{ij} is the number of such paths containing i and j. The purpose of the restriction $i>j$ is two-fold. The first is to eliminate repeated terms since $P_{ij_r}\Big|_{\substack{i=\lambda\\ j=\mu}} = P_{ij_r}\Big|_{\substack{i=\mu\\ j=\lambda}}$. The second reason is to eliminate the meaningless case of $i=j$. Physically, F_1 is the sum of all paths connecting C and C' containing edges in G. In the example of Fig. 5, $F=x_1x_7x_9+x_1x_4x_8x_9+x_2x_4x_7x_9+x_2x_8x_9$. F_1 may be written as a single summation; that is, $F_1=\sum_k^p P_{\hat{G}_k}$ where $P_{\hat{G}_k}$ is the k^{th} path connecting C and C' but containing edges in $\hat{G}$.

b) Form the function $F'_1=\sum_k^p P_{\hat{G}_k}\oplus(\hat{G}\cap P_{\hat{G}_k})$. Eliminate repeated and redundant terms.

In the example of Fig. 5, $F'_1=x_1x_9+x_2x_9$.

c) Form the function $F_1'''=\sum_k^p\{P_{\hat{G}_k}\oplus(\hat{G}\cap P_{G_k}\}\oplus P_{\hat{\hat{G}}_k}$

d) Form $F_3=F_1'''+F_2$, where all the terms in F_2 contain no edges of G.

e) Form $F_{\hat{G}\to\hat{\hat{G}}}$ by removing repeated and redundant terms.

Theorem 7. $F_{\hat{G}\to\hat{\hat{G}}}$ is the switching function of H where H is the graph obtained by replacing $\hat{G}$ by $\hat{\hat{G}}$ in graph G.

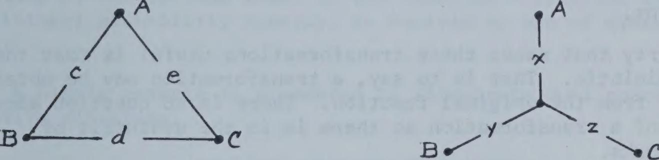
The usefulness of the substitution lies in its ability to take one SC function and transform it into a simpler SC function--simpler in the sense that the new function has fewer terms. What remains to be done is to formulate a method for finding appropriate subgraphs from a given SC function without having its realization at hand. There is one subgraph that is very easily detected, and when the appropriate substitution is made, a simpler function is guaranteed. This subgraph is a circuit with three edges or a delta. Substituting Y for the delta yields a simpler graph since there are two paths between any two vertices of a delta, while there is only one path between any two vertices of Y. The following simple theorem provides the key for detection of deltas in a realization.

Theorem 8. If the ring sum of any two terms of an SC function has exactly three elements in it, the ring sum corresponds to a delta.

Example 6: Given the function F=af+bdg+bcf+bceg+acdg+bdef.
Now, bceg⊕bdg=cde.

F_1=bdg+bcf+bceg+acdg+bdef, F_2=af.

The subgraphs under consideration are shown in Fig. 7.

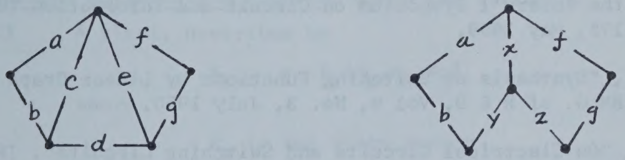


(a) : $\hat{G}$ (b): $\hat{\hat{G}}$

Figure 7

F_3=af+bYZg+bXYf+bYZg+aXZg+bXYf, $F_{\hat{G}\to\hat{\hat{G}}}$=af+bYZg+bXYf+zXZg.

The realization of these graphs are shown in Fig. 8.



(a) (b)

Figure 8

CONCLUSIONS

It is interesting to note that while any functional transformation of an SC function is guaranteed to be SC, there exists a class of functions that are not SC functions and which are mapped into SC functions by some functional transformation. Needless to say, these transformed functions pass all necessary tests for realizability. A simple example will illustrate the point.

Example 7: Consider the function F $F=x_1x_2x_4+x_2x_3x_5+x_1x_3x_6+x_4x_5x_6$. This function is known to be nonrealizable yet it satisfies the odd-ring-sum condition and the terminal-cutset condition. If the terminal transformation is applied with respect to x_1, the resulting function is

$$F(x_1)=x_1+x_2x_5x_6+x_3x_4x_5$$

and is realizable by graph G of Fig. 9.

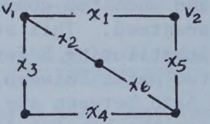


Figure 9

If, however, the path-cutset transformation is applied to F, the resulting function is F*.

$$F^*=x_1x_5+x_2x_6+x_3x_4+x_1x_2x_4+x_2x_3x_5+x_1x_3x_6+x_4x_5x_6$$

which does not satisfy the terminal cutset condition and, therefore, is not realizable. It should be pointed out that all functional transformations presented here preserve the odd-ring-sum condition. It would be nice to have a transformation that preserves the odd-ring-sum condition for SC functions but would not preserve it for some subset of non-SC functions.

The property that makes these transformations useful is that they are all deterministic. That is to say, a transformation may be obtained systematically from the original function. There is no question about the existence of a transformation as there is in the synthesis of switching networks.

REFERENCES

1. Gould, R., "Theory of Switching", Computational Lab., Progress Report No. BL-18, Harvard Univ., May 1957.

2. Lofgren, L., "Irredundant and Redundant Boolean Branch-Networks", Trans. of the Intern'l Symposium on Circuit and Information Theory, CT-6, 158-175, May 1959.

3. Mayeda, W., "Synthesis of Switching Functions by Linear Graph Theory", IBM J. of R & D, Vol 4, No. 3, July 1960.

4. Seshu, S., "On Electrical Circuits and Switching Circuits", IRE Trans. on Circuit Theory, CT-3, No. 3, 172-178, Sept. 1956.

5. Chan, S.P., "Application of Graph Theory to the Synthesis of Single-Contact Networks", to be published by the Journal of Franklin Institute in the November 1965 issue.

6. Chan, S.P., "Topological Analysis and Synthesis of Single-Contact Networks", Final Report, NSF Grant #GP-2760, July 1965.

COMBINED OPTIMAL CONTROL AND ESTIMATION*

LEWIS MEIER
Stanford Research Institute, Menlo Park, California

ABSTRACT

This paper presents an alternate formulation of Feldbaum's "dual-control" problem which was discussed in a series of four articles published in the Journal of Automation and Remote Control. The new formulation, referred to as the combined optimization problem, considers the determination of optimal control policies for a plant subjected to random disturbances and observed by means of a noisy measurement system. Adaptive problems, in which some of the parameters of the plant or random disturbances are incompletely known, can be considered under this formulation by augmenting the plant state vector. By use of this procedure the combined optimization problem is shown to be equivalent to the dual-control problem.

Two iterative equations are developed for solving the combined problem. Estimation, which is defined as the calculation of the conditional probability density of the state of the plant, given all available information, is governed by an iterative equation derived by application of Bayes' Rule. Control, which is defined as the determination of the optimum input to the plant as a function of the conditional probability density, is derived by use of dynamic programming.

A simple example is presented to illustrate dual control in a non-adaptive situation.

COMBINED OPTIMIZATION THEORY

Problem Statement

The combined optimization problem, which is illustrated in Fig. 1, is as follows:

Combined Optimization Problem. Given the following:

(1) A plant, described by

$$x_{k+1} = f_k(x_k, u_k, w_k) \tag{1}$$

where

x_k is the state vector

u_k is the input vector

w_k is the disturbance vector, assumed to be white

(2) A measurement system, described by

$$z_k = h_k(x_k, v_k) \tag{2}$$

---

*The research presented here was performed for NASA Ames Research Center as part of Contract NAS 2-2457.

where

z_k is the measurement vector

v_k is the measurement noise vector, assumed to be white

(3) The probability distributions

(a) $p(x_0)$ (3)

(b) $p(w_i)$ $i = 0,\ldots,N$ (4)

(c) $p(v_i)$ $i = 0,\ldots,N$ (5)

(4) The performance index

$$J = E\left\{\sum_{i=0}^{N} \ell(x_i, u_i, i)\right\} \tag{6}$$

(5) The admissibility constraint

$$u_i \in \Omega_i,$$

find the admissible controller that minimizes J, where:

(1) A controller is defined as any algorithm which at time k generates u_k as a function of the present and all past measurements $(z_k,\ldots,z_0)$.

(2) An admissible controller is defined as any controller which, when used in the closed-loop system shown in Fig. 1, yields admissible u_i.

Controller. The controller can be broken into two parts: The estimation system governed by

$$\hat{x}_{k+1} = \hat{f}\,(\hat{x}_k, z_k, k) \tag{7}$$

and the control law

$$u_k = \hat{h}\,(\hat{x}_k, z_k, k). \tag{8}$$

The quantity $\hat{x}_k$ is not in general what is thought of as the estimate of x_k. In particular it may have a dimension quite different from x_k. Any controller, optimal or suboptimal, may be represented in the form given by Eqs. (7) and (8); $\hat{x}_k$ is nothing more than the internal state of the controller in such a model.

Literature on the Problem

Several special cases of the combined optimization problem have been treated extensively in the past. If $z_k \equiv x_k$, then the problem reduces to the stochastic optimal control problem;[1*] and if, in addition,

---

* Wonham[11] and others have referred to the combined optimization problem defined above as the stochastic control problem; we prefer Bellman's more restrictive definition.[1]

$w_k \equiv 0$, then the problem reduces to the conventional optimal control problem. The optimal estimation problem results when f is independent of u; in this case u_k, the computer output, is the best estimate of x_k under criteria provided by J. Gunckel,[2] Tou and Joseph,[3] and Kalman[4] have solved the combined problem under the conditions that Eqs. (1) and (2) are linear; that $p(x_0), p(v_k)$, and $p(w_k)$ are Gaussian; and that $\ell(x_k, u_k, k)$ is quadratic.

Sussman[5] and Aoki[6] have considered the general problem as stated above. Feldbaum[7,8] treats the same problem in a different, but entirely equivalent formulation. (Meier[9] contains a demonstration that Feldbaum's dual-control problem and the combined optimization problems are the same.) Kushner[10] and Wonham[11] have done considerable work on the continuous time variant of this problem; and Refs. 10 and 11 contain bibliographies listing their work as well as related work by other authors.

The author has given a much more detailed version of the work presented here in an SRI report contracted by NASA (Ref. 9).

Simplified Derivation of Solution

In this section we present a simplified derivation of the solution of the combined optimization problem.

Stochastic Control Problem. Before the general problem is treated, we consider the solution of the special case in which the measurement system is perfect (i.e., $z_k = x_k$). For this case, no estimation is necessary; it is referred to as the Stochastic Control Problem. Bellman has derived a recursive equation for solution of this problem:[1] If $I(x_k, k)$ is defined by

$$I(x_k,\ k) = \min_{u_k(x_k),\ldots,u_N(x_N)} E\left\{\sum_{i=k}^{N} \ell(x_i, u_i, i)/x_k\right\} \tag{9}$$

then $u_k(x_k)$ is found by solution of

$$I(x_k, k) = \min_{u_k}\left[\ell(x_k, u_k, k) + \underset{w_k}{E}\left\{I(f_k(x_k, u_k, w_k), k+1)/x_k\right\}\right] \qquad k < N$$

$$I(x_N, N) = \min_{u_N} \ell(x_N, u_N, N) \tag{10}$$

where the w_k under the E indicates that the expectation is to be taken with respect to w_k.

Information State. In general, two sources of information about the state of the plant exist: the past control inputs and the past and present measurements. Thus, any prediction about the future behavior of the plant will be an expectation conditioned upon the values of these quantities. This expectation can be calculated from the state and measurement equations, the probability densities of the noises and the disturbances, and the conditional probability density of the plant state given the past control inputs and measurements. The latter quantity is the only one that is not known a priori;* hence it appears intuitive that the control can be computed as a function of this quantity.

In Ref. 9 it is shown that the optimal control input can indeed be calculated as a function of the conditional probability density of state. The $\hat{x}$ appearing in Eqs. (7) and (8) is the conditional probability density: Estimation is calculation of the conditional probability density of the state, and control is calculation of the control input as a function of this conditional density.

Because it summarizes all past information and because it plays a role similar to the state of the plant in deterministic optimal control theory, the conditional probability density of the state will be referred to as the information state $\wp_k$.

Control Equation. In the next section we show that $\wp_k$ satisfies the recursion relation

$$\wp_{k+1} = F_k\ (\wp_k, u_k, z_{k+1}). \tag{11}$$

Equation (11) has the form of a state equation for a plant with state $\wp_k$, input u_k, and disturbance z_{k+1}.

Consider the performance index in Eq. (6). By use of a well known identity on conditional expectations[12] each term of the summation may be written

$$E\{\ell(x_i, u_i, i)\} = E\ \underset{x_i}{\{}\ell(x_i, u_i, i)/\wp_i\}$$

$$= E\ \{L(\wp_i, u_i, i)\} \tag{12}$$

---

*If the system equations or probability densities contain unknown parameters, the state can be augmented to include these parameters.

where

$$L(\mathcal{P}_i,u_i,i) \overset{\Delta}{=} \underset{x_i}{E} \{\ell(x_i,u_i,i)/\mathcal{P}_i\} = E\{\ell(x_i,u_i,i)/z_i,u_{i-1}\} \tag{13}$$

and the capitals are used to denote past history of the variable $(Z_k = z_k, \ldots, z_0)$.

Hence we may rewrite Eq. (13) as

$$J = E\{\sum_{i=0}^{N} L(\mathcal{P}_i,u_i,i)\} \tag{14}$$

Comparing Eqs. (11) and (14) with Eqs. (1) and (6) we see that the former two equations define a stochastic control problem in which $\mathcal{P}_k$ is the state, z_{k+1} is the disturbance, and u_k is the control. This problem differs from the ordinary stochastic control problem in that the state is not finite dimensional; however, in the derivation of Eq.(10) no property of the dimension of x_i is used. A second difference is that the disturbance is not white; however, it is not necessary that the disturbance be white -- it is only necessary that $\mathcal{P}_k$ be a Markov process, which it is. Hence, we may write at once:

$$I^*(\mathcal{P}_k,k) = \min_{u_k}\left[L(\mathcal{P}_k,u_k,k) + \underset{z_{k+1}}{E}\{I^*(F_k(\mathcal{P}_k,u_k,z_{k+1}),\ k+1)/\mathcal{P}_k\}\right] \quad k<N$$

$$I^*(\mathcal{P}_N,N) = \min_{u_N} L(\mathcal{P}_N,u_N,N) \tag{15}$$

where

$$I^*(\mathcal{P}_k,k) \overset{\Delta}{=} \min_{u(\mathcal{P}_k),\ldots,u(\mathcal{P}_N)} E\left\{\sum_{i=k}^{N} L(\mathcal{P}_i,u_i,i)/\mathcal{P}_k\right\}. \tag{16}$$

which is identical with Eq. (2-32) of Ref. 9. Because solution of Eq.(15) enables one to determine the optimum control law (i.e., to determine u_k as a function of $\mathcal{P}_k$), we will refer to it as the control equation.

Estimation Equation. In this section we present the form of the F given in Eq. (11); this equation will be referred to as the estimation equation. It may be used to update the probability density in real time or alternatively we may use Eq. (11) in conjunction with Eq. (15) to determine the controls u_k as a function of z_k rather than $\mathcal{P}_k$. These two methods of specifying control are treated in detail in Ref. 9.

An equation for updating the conditional probability of the state of a plant with no control inputs is derived by application of Bayes'

Rule, by Peschon.[13] This equation may be modified to handle control inputs by simply replacing $p(x_{k+1}/x_k)$ by $p(x_{k+1}/x_k, u_k)$. The result is

$$p(x_{k+1}/Z_{k+1}, U_k) = \frac{p(z_{k+1}/x_{k+1}) \int_{x_k} p(x_{k+1}/x_k, u_k)\, p(x_k/Z_k, U_{k-1})\, dx_k}{\int_{x_{k+1}} p(z_{k+1}/x_{k+1}) \int_{x_k} p(x_{k+1}/x_k, u_k)\, p(x_k/Z_k, \underset{0<k}{U_{k-1}})\, dx_k\, dx_{k+1}}$$

$$p(x_o/Z_o, U_{-1}) = \frac{p(z_o/x_o)\, p(x_o)}{\int_{x_o} p(z_o/x_o) p(x_o)\, dx_o} \qquad (17)$$

Since $\wp_k \triangleq p(x_k/Z_k, U_{k-1})$, Eq. (17) indeed specifies the form of F in Eq. (11).

Comments. Because the combined optimization problem is equivalent to an infinite dimensional control problem, we can expect considerable difficulties in calculating exact solutions in general. Suitable approximations are proposed in Ref. 9; investigation of approximations is a major area of unfinished research on even the classical optimal control problem, not to mention the combined optimization problem.

In two special cases, however, the combined optimization problem is not infinite dimensional: If Eqs. (1) and (2) are linear; if Eqs. (3), (4), and (5) are Gaussian; and if Eq. (6) is quadratic; then Eqs. (15) and (17) reduce to a well known matrix equation. If x_k can take only a finite number of values, then $\wp_k$ is just the vector of probabilities that the plant is in one of its possible states and Eqs. (15) and (17) become finite sets. The solution to the linear case is summarized by Peschon *et al*;[14] Ref. 9 presents the detailed derivation of the results. The second half of this paper presents an example of a finite state system. Solution of Eq. (17) requires knowledge of $p(x_{k+1}/x_k, u_k)$ and $p(z_k/x_k)$. The first of these probability densities can be obtained from the state equation and $p(w_k)$; $p(x_{k+1}/x_k, u_k)$ is an alternative way of describing a randomly disturbed plant. Similarly, $p(z_k/x_k)$ can be obtained from knowledge of the measurement equation and it is an alternative way of describing the measurement system.

DUAL CONTROL THEORY

Definition

The purpose of feedback in a control system is to gain information about the plant.[14] In addition to the use of sensors to gather information, and filters to process information, there is a third method of gaining usable information about the state of the plant -- control action. Intuitively the concept of using control action to gain information involves nothing more than using test signals as inputs in order to gain information; hence the control input can be used to gain information as well as control the plant. Since the input that provides the most expected information is not necessarily the one that is most likely to bring the plant to a desired state, it is necessary to consider the compromise between the control and informational uses of the control input. In recognition of the dual purpose of control inputs, Feldbaum[7,8] refers to the problem as Dual Control.

The most obvious application of dual control theory is to adaptive control systems--indeed this was the original motivation behind Feldbaum's work; however, control action can be used to help estimate the state of a plant in a non-adaptive situation, as the example illustrates.

Example

To illustrate the concept of dual control, the following simple example is presented. This system is finite state with x_k, u_k, v_k, and z_k being binary variables which can take the values 0 or 1. It is described by

$$x_{k+1} = x_k \oplus u_k \tag{18}$$

$$z_k = x_k \cdot v_k \tag{19}$$

where the logic functions $\oplus$ and $\cdot$ are defined in Table 1.

Table 1

LOGIC FUNCTIONS

$x_k \backslash u_k$	0	1
0	0	1
1	1	0

Exclusive or ($\oplus$)

$x_k \backslash v_k$	0	1
0	0	0
1	0	1

And ($\cdot$)

The noise v_k takes the value 1 with probability $\frac{1}{2}$ independently.

In this example the conditional probability distribution can be described by two numbers: the conditional probability that x_k is zero and the conditional probability that x_k is one. Since these two numbers must add up to one, a suitable information state is

$$p_k \overset{\Delta}{=} \Pr(x_k = 1/z_k, \ldots, z_o, u_{k-1}, \ldots, u_o). \tag{20}$$

Because of the discrete nature of this problem, the integrations in the estimation equation must be changed to summations. The discrete-state combined optimization problem was solved independently by Aström.[15]

The entropy $H(p_k, k)$ of the information state is defined by

$$H(p_k, k) = p_k \log_2 p_k + (1 - p_k) \log_2(1 - p_k). \tag{21}$$

The expected entropy $U(u_k, p_k, k)$ at the next time when control u_k is used is plotted along with $H(p_k, k)$ in Fig. 2. From this figure, it may be seen that most information is gained when $u_k = 0$ is used for $p_k < \frac{1}{2}$, and $u_k = 1$ for $p_k > \frac{1}{2}$.

The control that is most likely to take the system to zero is the one that provides most information; hence, if the desired state is zero there is no conflict between the uses of control action. On the other hand, if the desired state is one, then the control that provides most information differs from that which is most likely to take the system to the desired state.

We take N = 2 and consider two cost functions that differ only in the desired state. For desired state zero, the cost function is

$$\begin{aligned} \ell(u_k, u_k, k) &= x_k & k < 2 \\ &= 10x_k & k = 2 . \end{aligned} \tag{22}$$

The optimal control policy for this problem is given in Table 2 and the return functions $I^*(p_k, k)$ are presented in Fig. 3. For desired state one, the cost function is

$$\begin{aligned} \ell(u_k, x_k, k) &= x_k \oplus 1 & k < 2 \\ & 10(x_k \oplus 1) & k = 2 . \end{aligned} \tag{23}$$

Table 2 is the optimal policy for this case, and Fig. 4 presents the return functions with reverse scale for ease of comparison with Fig. 3.

Table 2

OPTIMAL POLICY FOR DESIRED STATE = ZERO

k	0	1	2
$u_k^0\ (p_k)$	$0\ \ p_o < \frac{1}{2}$ $1\ \ p_o > \frac{1}{2}$	$0\ \ p_1 < \frac{1}{2}$ $1\ \ p_1 > \frac{1}{2}$	$0\ \ p_2 < \frac{1}{2}$ $1\ \ p_2 > \frac{1}{2}$
$I^*(p_k,k)$	$7p_o, p_o < \frac{1}{2}$ $p_o+6(1-p_o), p_o > \frac{1}{2}$	$11p_1, p_1 < \frac{1}{2}$ $p_1=10(1-p_1), p_1 > \frac{1}{2}$	$10p_2$

Table 3

OPTIMAL POLICY FOR DESIRED STATE = ONE

k	0	1	2
$u_k^0\ \ (p_k)$	$1\ \ p_o < 1/7$ $0\ \ 1/7 < p_o < \frac{1}{2}$ $1\ \ \frac{1}{2} < p_o < 6/7$ $0\ \ 6/7 < p_o$	$1\ \ p_1 < \frac{1}{2}$ $0\ \ p_1 > \frac{1}{2}$	$1\ \ p_2 < \frac{1}{2}$ $0\ \ p_2 > \frac{1}{2}$
$I^*(p_k,k)$	$1+10p_o\ \ \ p_o < 1/7$ $2+3p_o\ \ \ 1/7 < p_o < \frac{1}{2}$ $6-5p_o\ \ \ \frac{1}{2} < p_o < 6/7$ $12(1-p_o)\ \ \ 6/7 < p_o$	$1+9p_1,\ p_1 < \frac{1}{2}$ $11(1-p_1), p_1 > \frac{1}{2}$	$10(1-p)$

The cost of being in the wrong state at Time 2 is very high; therefore it is important that the proper control be made at Time 1. Table 3 indicates that when the desired state is one and the knowledge of the state at Time 0 is poor (i.e., p_0 is close to $\frac{1}{2}$) then the optimal control is the one that provides most information, thus insuring that the probability of taking the proper control at Time 1 is highest.

Based upon this example we can conclude that there exist cases where the input that provides most information differs from the input that is optimal from an immediate control point of view. In these cases the optimal policy is a compromise between gaining information and bringing the plant to its desired state. In general when such a compromise has to be made the cost is higher than for analogous cases where the compromise is not necessary.

References

1. R. Bellman, Adaptive Control Processes: A Guided Tour, p. 156 (Princeton University Press, 1961).
2. T. L. Gunckel, "Optimum Design of Sampled-Data Systems with Random Parameters," S.E.L. T.R. No. 2102-2, Stanford Electronics Laboratory, Stanford, California (April 1961).
3. P. D. Joseph and J. T. Tou, "On Linear Control Theory," Trans. AIEE, Part II, 80, p. 18 (1961).
4. R. E. Kalman, "On the General Theory of Control Systems," Proc. of the First IFAC Congress, Moscow, pp. 481-492 (1960).
5. R. Sussman, "Optimal Control of Systems with Stochastic Disturbances," Series 63, No. 20, Electronics Research Laboratory, University of California, Berkeley, California (November 1963).
6. M. Aoki, "Optimal Bayesian and Min-Max Controls of a Class of Stochastic and Adaptive Dynamic Systems," 1965 IFAC Tokyo Symposium of Systems Engineering for Control System Design, Preprints, pp. II-21 to II-31.
7. A. A. Feldbaum, "Optimal Systems," Chapter from Disciplines and Techniques of System Control, John Peschon (Blaisdell, 1965).
8. A. A. Feldbaum, "Dual Control Theory, I," Automatika i Telemekhanika (September, 1960).
9. L. Meier, "Combined Optimal Control and Estimation Theory," NASA Contractor Report by SRI for Ames Research Center, Contract NAS 2-2457 (October 1965).
10. Harold J. Kushner, "Some Problems and Recent Results in Stochastic Control," IEEE National Convention Record, Part 6 (1965).
11. W. M. Wonham, "Stochastic Problems in Optimal Control," IEEE National Convention Record (1963).
12. E. Parzen, Modern Probability Theory and its Applications, p. 384 (John Wiley & Sons, Inc., 1960).
13. J. Peschon, "Review and Extension of Real-Time Estimation Techniques" 1966 IFAC paper, available on request.
14. J. Peschon, L. Meier, R. E. Larson, W. H. Foy, and C. H. Dawson, "Information Requirement for Guidance and Control Systems," Final Technical Report on Contract NAS 2-2457 with NASA Ames Research Center, to be published.
15. K. J. Åström, "Optimal Control of Markov Processes with Incomplete State Information," TP 18-137, IBM Nordic Laboratory, Sweden (July 1964).

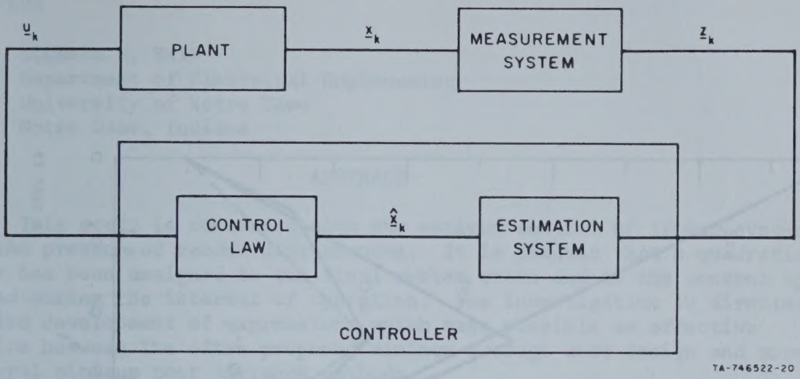

$\underline{u}_k$
PLANT
$\underline{x}_k$
MEASUREMENT SYSTEM
$\underline{z}_k$
CONTROL LAW
$\hat{\underline{x}}_k$
ESTIMATION SYSTEM
CONTROLLER
TA-746522-20

FIG. 1 COMBINED OPTIMIZATION PROBLEM

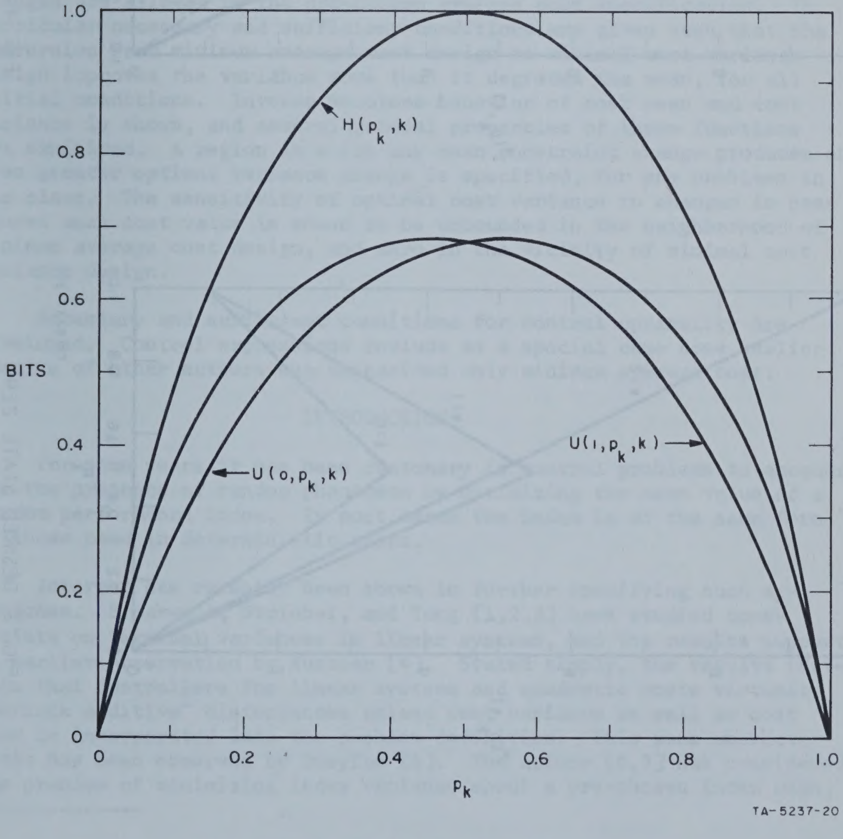
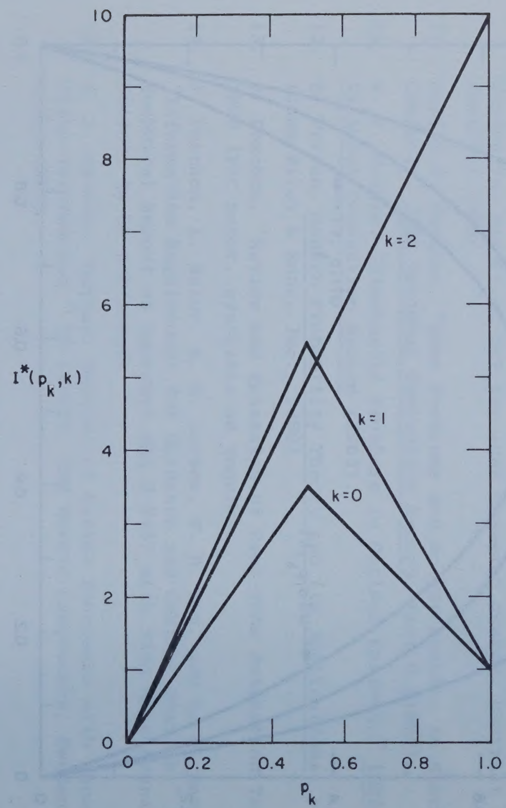
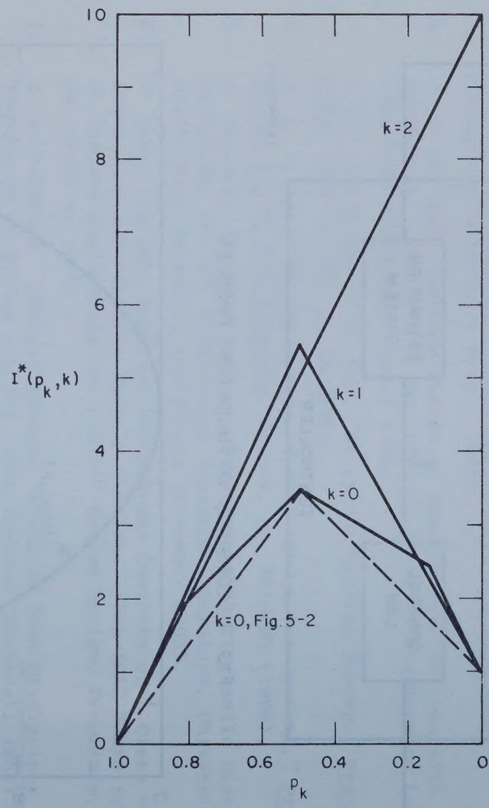


FIG. 2 EXPECTED ENTROPY



TA-5237-21

FIG. 3 DESIRED STATE ZERO



TA-5237-23

FIG. 4 DESIRED STATE ONE

RELATIVE COSTS OF MEAN AND VARIANCE CONTROL FOR A CLASS OF LINEAR, NOISY SYSTEMS

MICHAEL K. SAIN
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

ABSTRACT

This study is concerned with the optimum control of linear systems in the presence of random disturbances. It is assumed that a quadratic cost has been assigned to the final system state and to the control applied during the interval of operation. The investigation is directed toward development of expressions which make possible an effective choice between the often proposed minimum average cost design and more general minimum cost variance methods.

The main results are incremental and differential representations for changes in optimal cost variance occurring when corresponding changes are allowed in the pre-chosen average cost specification. In particular necessary and sufficient conditions are given such that the conversion from minimum average cost design to minimal cost variance design improves the variance more than it degrades the mean, for all initial conditions. Inverse monotone behavior of cost mean and cost variance is shown, and several general properties of these functions are exhibited. A region in which any mean constraint change produces an even greater optimal variance change is specified, for any problems in the class. The sensitivity of optimal cost variance to changes in pre-chosen mean cost value is shown to be unbounded in the neighborhood of minimum average cost design, and zero in the vicinity of minimal cost variance design.

Necessary and sufficient conditions for control optimality are developed. Control expressions include as a special case some earlier results of other authors who emphasized only minimum average cost.

INTRODUCTION

For some years it has been customary in control problems to account for the presence of random phenomena by optimizing the mean value of a random performance index. In most cases the index is of the same form as those used in deterministic cases.

Interest has recently been shown in further specifying such approaches. Breakwell, Striebel, and Tung [1,2,3] have studied constraints on terminal variances in linear systems, and the results support an earlier observation by Kushner [4]. Stated simply, the results indicate that controllers for linear systems and quadratic costs virtually overlook additive[1] disturbances unless cost variance as well as cost mean is incorporated into the problem definition. This same characteristic has been observed by Dreyfus [5]. The author [6,7] has considered the problem of minimizing index variance about a pre-chosen index mean, for linear systems and quadratic costs. The latter work suggested the present investigation.

---

[1] The random functions appear as linear additions to the plant equations.

The importance of this work is to give conditions under which a minimal variance design offers decided improvement over the minimum average cost design.

PROBLEM STATEMENT AND PRINCIPAL RESULTS

The framework of the paper is simply stated in the following manner. Find the control $\underline{u}_\mu^o$ which minimizes

$$V_\mu\{J(\underline{u})\} = E\{[J(\underline{u}) - E\{J(\underline{u})\}]^2\} \tag{1}$$

subject to constraints

$$E\{J(\underline{u})\} = M_\mu \quad ; \tag{2}$$

$$J(\underline{u}) = \int_{t_0}^{t_F} ||\underline{u}||_R^2 \, d\tau + ||\underline{x}(t_F)||_P^2 \quad ; \tag{3}$$

$$\dot{\underline{x}} = A(t)\underline{x} + B(t)\underline{u} + C(t)\underline{v}, \quad \underline{x}(t_0) = \underline{x}_0 \quad . \tag{4}$$

Here V_μ represents[2] the cost variance and M_μ the cost mean associated with (3); the ensemble average $E\{\}$ is conditioned upon knowledge of the $\underline{v}$ process prior to time t_0. The vectors $\underline{x}$, $\underline{u}$, and $\underline{v}$ are defined in n, m, and r-dimensional real spaces; $||\underline{a}||_E^2$ denotes the squared Euclidean norm[3] $\underline{a}'E\underline{a}$; A, B, C, P and R are conformal matrices, the last two being symmetric, nonnegative and positive definite respectively.

It is convenient to retain the Lagrange multiplier μ as a parameter. The case $\mu \to \infty$ is called the minimal mean problem; the case $\mu \to 0^+$ is termed the minimal variance problem. The main results:

I. Parametric Slope

Pre-chosen average cost M_μ (optimal cost variance V_μ^o) is a monotonically decreasing (increasing) function of μ for $\mu > 0$. This result ensures that any compromise which can be made with minimum average cost design will in fact result in a smaller cost variance.

II. Incremental Representation

Let the constraint (2) be changed from M_{μ_1} to M_{μ_2}; then the change in optimal variance is

$$V^o_{\mu_2} - V^o_{\mu_1} = - \int_{\mu_1}^{\mu_2} s \, dM_s \tag{5a}$$

for $\mu_1, \mu_2 > 0$. The change in constraint which is necessary to balance cost variance adjustment is found by evaluating

---

[2] The subscript μ refers to a Lagrange multiplier introduced subsequently in connection with the constraint (2).

[3] Superscript ' denotes transposition.

$$M_{\mu_2} - M_{\mu_1} = - \int_{\mu_1}^{\mu_2} \frac{1}{s} dV_s^o \; . \tag{5b}$$

Related to (5) are the observations:

A. Incremental Ratio Estimation. Because of the monotone character of these integrator functions, the ratio of change in optimal variance to change in mean value can be bounded; namely

$$\mu_1 \le \frac{V^o_{\mu_2} - V^o_{\mu_1}}{M_{\mu_1} - M_{\mu_2}} \le \mu_2 \; . \tag{6}$$

This expressions exhibits the point $\mu=1$ on the parametric half-line as critical in relative variance and mean considerations.

B. Differential Relations. Comparative behavior of cost variance and cost mean can also be stated differentially. First, directly--

$$\frac{dV^o_\mu}{dM_\mu} = - \mu \; ; \tag{7a}$$

and second, parametrically--

$$\frac{dV^o_\mu}{d\mu} = -\mu \frac{dM_\mu}{d\mu} \; . \tag{7b}$$

Again the point $\mu=1$ is shown to be of great interest.

C. Incremental Sensitivity. For a given differential constraint change, the resulting differential optimal variance change is greatest in the neighborhood of the minimum average cost design and least in the neighborhood of the minimum cost variance design.

III. Relative Design Costs

From the statements above, it is clear that relaxation of average cost constraints will effect an improvement in cost variance. The comparison here involves the maximum possible variance improvement and the corresponding increase in average cost. The difference

$$(V^o_\infty - V^o_0) - (M_0 - M_\infty) \ge 0 \tag{8}$$

for all initial conditions $\underline{x}_0$ if and only if the matrix[4]

---

[4] In the limit $\mu \to 0^+$ both P^{-1} and $Q^{-1}(t_F)$ are required to exist; the former condition is simply a strengthened assumption. Existence of $\underline{H}^{-1}(t_F)$ and $Q^{-1}(t_F)$ are associated with the controllability of $\dot{\underline{\theta}}_1 = A(t)\underline{\theta}_1 + B(t)\underline{u}_1$ and $\dot{\underline{\theta}}_2 = A(t)\underline{\theta}_2 + C(t)\underline{u}_2$ respectively; in the latter case $S(\tau,\sigma)$ is assumed nonsingular.

$$4PQ(t_F)P - P - H^{-1}(t_F) \tag{9}$$

is nonnegative definite. The matrices $Q(t_F)$ and $H(t_F)$ are defined by

$$Q(t_F) = \int_{t_0}^{t_F}\int_{t_0}^{t_F} \Phi(t_F, \tau)C(\tau)S(\tau,\sigma)C'(\sigma)\Phi'(t_F,\sigma)d\sigma d\tau \tag{10}$$

where $S(\tau,\sigma)$ is the $\underline{v}$ process covariance matrix, and

$$H(t_F) = \int_{t_0}^{t_F} \Phi(t_F, \tau)B(\tau)R^{-1}(\tau)B'(\tau)\Phi'(t_F, \tau)d\tau ; \tag{11}$$

$\Phi(\cdot,\cdot)$ is the state transition matrix corresponding to (4).

IV. Geometrical Properties

The shape and properties of typical variance-mean curves are shown in Figure 1 for parametric and Figure 2 for direct representations.

ANCILLARY RESULTS

The following auxiliary facts have contributed to the principal conclusions delineated above.

V. Optimum Control

The solution to the problem posed in (1-4) is

$$\begin{aligned}\underline{u}^o_\mu(t) = {} & -K(t_F,t)T_\mu(t_F)[\Phi(t_F,t_0)\underline{x}_0 + \underline{m}_w(t_F)] \\ & - \tfrac{2}{\mu}K(t_F,t)L_\mu^{-1}(t_F)P\underline{q}(t_F)\end{aligned} \tag{12}$$

with the definitions

$$K(t_F,t) = R^{-1}(t)B'(t)\Phi'(t_F,t) \quad ; \tag{13}$$

$$G_\mu(t_F) = P(I + \tfrac{4}{\mu}Q(t_F)P) \quad ; \tag{14}$$

$$L_\mu(t_F) = I + G_\mu(t_F)H(t_F) \quad ; \tag{15}$$

$$T_\mu(t_F) = G_\mu(t_F)[L_\mu'(t_F)]^{-1} \quad ; \tag{16}$$

$$\underline{w}(t) = \int_{t_0}^{t} \Phi(t, \tau)C(\tau)\underline{v}(\tau)d\tau \quad ; \tag{17}$$

$$\underline{m}_w = E\{\underline{w}\} \quad ; \quad \tilde{\underline{w}} = \underline{w} - \underline{m}_w \quad ; \tag{18}$$

$$v_{wp} = E\{||\tilde{\underline{w}}||_P^2\} \quad ; \quad \tilde{v} = ||\tilde{\underline{w}}||_P^2 - v_{wp} \quad ; \tag{19}$$

$$\underline{q} = E\{\tilde{v}\tilde{\underline{w}}\} \quad . \tag{20}$$

Expression (12) defines the unique, optimal control if and only if $H^{-1}(t_F)$ exists, under the sufficiency assumption $\mu \geq 0$. The special case of (12), namely $\underline{u}^o_\infty$, has been previously established [8].

VI. Optimal Cost Variance

The value of (1) associated with (12) is

$$V^{o}_{\mu} = q_s(t_F) + 4\underline{q}'(t_F)\underline{d}_{\mu}(t_F) + 4||\underline{d}_{\mu}(t_F)||^2_{Q(t_F)} \tag{21}$$

where

$$q_s = E\{\tilde{v}^2\} \; ; \tag{22}$$

$$\underline{d}_{\mu}(t_F) = [L'_{\mu}(t_F)]^{-1}[\Phi(t_F,t_0)\underline{x}_0 + \underline{m}_w(t_F) - \frac{2}{\mu}H(t_F)P\underline{q}(t_F)]. \tag{23}$$

One special case is of primary interest, namely

$$V^{o}_{0} = V^{o}_{min} = q_s(t_F) - ||\underline{q}(t_F)||^2_{Q^{-1}(t_F)} \; ; \tag{24}$$

the second term in this expression clearly shows the improved performance which is possible when the disturbance process is correlated. In particular, $q(t_F)$ vanishes if $\underline{v}$ is a white, gaussian process with mutually uncorrelated components.

VII. Mean Constraint Function

By the monotone results of I, M_{μ} depends on μ in a one-one manner for $\mu > 0$.

$$M_{\mu} = ||\underline{d}_{\mu}(t_F)||^2_P + ||G_{\mu}(t_F)\underline{d}_{\mu}(t_F) + \frac{2}{\mu}P\underline{q}(t_F)||^2_{H(t_F)} + v_{wp}(t_F). \tag{25}$$

Constraining M_{μ} or μ is then appropriate.

METHOD OF SOLUTION

By linearity, $\underline{x}$ is the sum of a deterministic component $\underline{z}$ and a random component $\underline{w}$, where

$$\dot{\underline{z}} = A(t)\underline{z} + B(t)\underline{u} \; , \; \underline{z}(t_0) = \underline{x}_0. \tag{26}$$

The cost function (1) becomes

$$V_{\mu} = q_s(t_F) + 4\underline{q}'(t_F)P[\underline{z}(t_F) + \underline{m}_w(t_F)] + 4||\underline{z}(t_F) + \underline{m}_w(t_F)||^2_{PQ(t_F)P} \; ; \tag{1'}$$

and the constraint (2) is given by

$$v_{wp}(t_F) + ||\underline{z}(t_F) + \underline{m}_w(t_F)||^2_P + \int_{t_0}^{t_F} ||\underline{u}||^2_R d\,\tau = M_{\mu} \quad . \tag{2'}$$

An equivalent problem therefore is to minimize

$$||\underline{z}(t_F) + \underline{m}_w(t_F)||^2_{P(\mu I + 4Q(t_F)P)} + 4\underline{q}'(t_F)P[\underline{z}(t_F) + \underline{m}_w(t_F)]$$

$$+ q_s(t_F) + [v_{wp}(t_F) + \int_{t_0}^{t_F} ||\underline{u}||^2_R d\,\tau]\;\mu \tag{1''}$$

subject to (26).

The solution to (1'') is well known; following [9,10] leads necessarily to[5] the Euler-Lagrange equations

$$\dot{\underline{\lambda}} + A'(t)\underline{\lambda} = \underline{0} \quad ; \tag{27}$$

$$\underline{\lambda}(t_F) = 4P\underline{q}(t_F) + 2P(\mu I + 4Q(t_F)P)(\underline{z}(t_F) + \underline{m}_w(t_F)) \ ; \tag{28}$$

and the stationarity condition

$$-2\mu R\underline{u} = B'(t)\underline{\lambda} \quad . \tag{29}$$

From the accessory minimum problem, the conditions $\mu > 0$ and the existence of $H^{-1}(t_F)$ are sufficient to ensure optimality and uniqueness. The latter inverse must necessarily exist in case $\mu \geq 0$.

The remaining steps are essentially algebraic in nature, and their details are omitted. Selected comments appear in the Appendix, however.

CONCLUSIONS

It has been shown, for the class of problems (1-4), that

1. Any relaxation in the usual minimum average cost design requirement will assuredly result in a decrease of cost variance.
2. In any such problem, the cost variance reduction always exceeds the increased average cost, up to a certain fixed point (μ=1).
3. If a minimum average cost problem is converted into a minimum cost variance problem, the total decrease in variance will exceed the total increase in average value for all initial conditions if and only if the matrix (9) is nonnegative definite.
4. The incremental sensitivity of cost variance to cost mean is greatest in the neighborhood of the minimum average cost design.
5. Incremental changes in cost variance and cost mean obey the representation (5).

These results clearly delineate conditions under which minimum average cost solutions may be unsatisfactory.

REFERENCES

1. F. Tung, "Linear Control Theory Applied to Interplanetary Guidance," IEEE Trans. on Automatic Control, Vol. AC-9, pp. 82-89; January, 1964.

2. J. V. Breakwell and F. Tung, "Minimum Effort Control of Several Terminal Components," SIAM J. on Control, Vol. 2, No. 3, pp. 295-316; 1964.

3. J. V. Breakwell and C. T. Striebel, "Minimum Effort Control in Interplanetary Guidance," Preprint 63-80, IAS Meeting, New York, N.Y.; 1963.

4. H. J. Kushner, "Near Optimal Control in the Presence of Small Stochastic Perturbations," Research Institute for Advanced Studies, Baltimore, Md., Tech. Rept. No. 63-22; April, 1964.

5. S. E. Dreyfus, "Some Types of Optimal Control of Stochastic Systems," SIAM J. on Control, Vol. 2, No. 1., pp. 120-134; 1964.

---

[5] Lagrange multiplier $\underline{\lambda}$ corresponds to (26).

6. M. K. Sain, "On Minimal-Variance Control of Linear Systems with Quadratic Loss," Ph.D. Thesis, U. of Illinois, 1965; also Tech. Rept. No. R-240, Coordinated Science Lab., U. of Illinois, Urbana, Ill; January, 1965.

7. M. K. Sain, "Control of Linear Systems According to the Minimal-Variance Criterion - A New Approach to the Disturbance Problem," Dept. of Electrical Engineering, U. of Notre Dame, Notre Dame, Ind., Tech. Report No. EE-656; June 8, 1965.

8. F. B. Tuteur and J. S. Tyler, Jr., "Optimal-Control Theory Applied to a Probabilistic Intercept Problem," *IEEE Trans. on Automatic Control*, Vol. AC-9, pp. 498-507; October, 1964.

9. J. V. Breakwell and Y. C. Ho., "On the Conjugate Point Condition for the Control Problem," *International J. of Engineering Science*, Vol. 2, No. 6, pp. 565-579; March, 1965.

10. S. Baron, A. E. Bryson, Jr., and Y. C. Ho, "Differential Games and Optimal Pursuit-Evasion Strategies," JACC Preprints, pp. 37-40; June, 1965.

APPENDIX

Numerous examples of the solution of (27-29) are available in the literature; the procedure will be indicated here only in outline form.

1. Express $\underline{\lambda}(t)$ in terms of $\underline{\lambda}(t_F)$.

2. Using (28-29), express $\underline{u}$ in terms of $\underline{z}(t_F)$.

3. Substitute the last result into the defining equations for $\underline{z}(t_F)$ and solve for $\underline{z}(t_F)$.

4. The $\underline{u}$ relation of step 2 is then defined completely.

Once the optimal control is established, M_μ and V_μ^o may be found directly. Limiting cases leading to (9) present no special difficulty if terms are grouped properly.

The result (9) is, however, hastened by writing the appropriate expressions in similar forms. It is trivial to state that

$$M_\infty = ||\underline{d}_\infty(t_F)||^2_{T_\infty(t_F)} + v_{wp}(t_F) \quad ; \tag{30}$$

$$\begin{aligned} M_0 = {} & \tfrac{1}{4}||P^{-1}Q^{-1}(t_F)\underline{q}(t_F)||^2_P + v_{wp}(t_F) \\ & + \tfrac{1}{4}||\Phi(t_F,t_0)\underline{x}_0 + \underline{m}_w(t_F) + P^{-1}Q^{-1}(t_F)\underline{q}(t_F)||^2_{H^{-1}(t_F)} \ . \end{aligned} \tag{31}$$

But for useful comparison these may be rewritten as

$$M_\infty = v_{wp}(t_F) - \tfrac{1}{4}||\underline{e}(t_F)||^2_{T_\infty(t_F)} - \underline{e}'(t_F)PL'_\infty(t_F)\underline{d}_\infty(t_F)$$

$$+ \frac{1}{4}\left|\left|2\underline{d}_{\infty}(t_F) + \underline{e}(t_F)\right|\right|^2_{T\infty(t_F)} \quad ; \tag{30)'}$$

$$M_0 = v_{wp}(t_F) - \frac{1}{4}\left|\left|\underline{e}(t_F)\right|\right|^2_{T_{\infty}(t_F)} - \underline{e}'(t_F)PL_{\infty}'(t_F)\underline{d}_{\infty}(t_F)$$

$$+ \frac{1}{4}\left|\left|2\underline{d}_{\infty}(t_F) + \underline{e}(t_F)\right|\right|^2_{L_{\infty}(t_F)H^{-1}(t_F)L_{\infty}'(t_F)} \tag{31)'}$$

where

$$\underline{e}(t_F) = P^{-1}Q^{-1}(t_F)\underline{q}(t_F). \tag{32}$$

In an analogous fashion, two cost variance pairs may be established; the second pair is

$$V_{\infty}^{o} = q_s(t_F) - \left|\left|\underline{q}(t_F)\right|\right|^2_{Q^{-1}(t_F)} + \left|\left|2\underline{d}_{\infty}(t_F) + \underline{e}(t_F)\right|\right|^2_{PQ(t_F)P} ; \tag{33}$$

$$V_0^{o} = q_s(t_F) - \left|\left|\underline{q}(t_F)\right|\right|^2_{Q^{-1}(t_F)} \quad . \tag{34}$$

The condition (9) follows straightforwardly.

The representation results may be demonstrated by differentiation and comparison for the functions V_{μ}^{o} and M_{μ} . Define

$$\underline{f}_{\mu}(t_F) = [2PQ(t_F)P\underline{d}_{\mu}(t_F) + P\underline{q}(t_F)] \quad ; \tag{35}$$

differentiation with respect to the parameter μ yields

$$\frac{dV_{\mu}^{o}}{d\mu} = \frac{8}{\mu^2}\left|\left|\underline{f}_{\mu}(t_F)\right|\right|^2_{H(t_F)L_{\mu}^{-1}(t_F)} \quad ; \tag{36}$$

$$\frac{dM_{\mu}}{d\mu} = \frac{-8}{\mu^3}\left|\left|\underline{f}_{\mu}(t_F)\right|\right|^2_{H(t_F)L_{\mu}^{-1}(t_F)} ; \tag{37}$$

Other results follow in like manner.

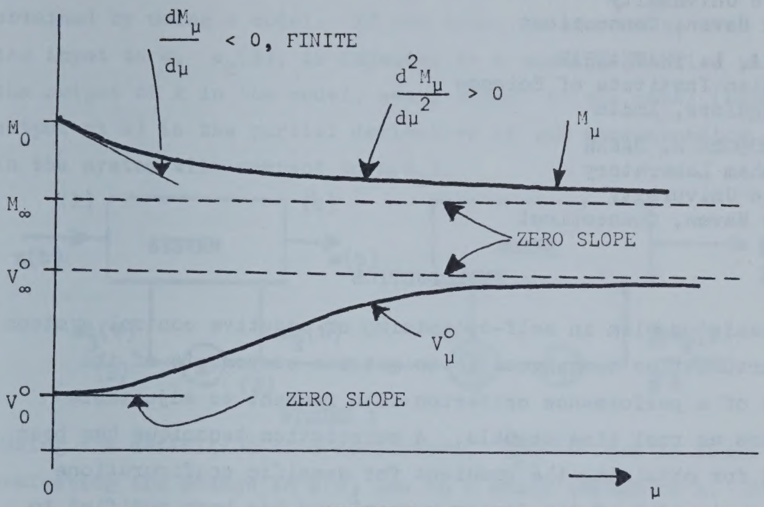


FIGURE 1

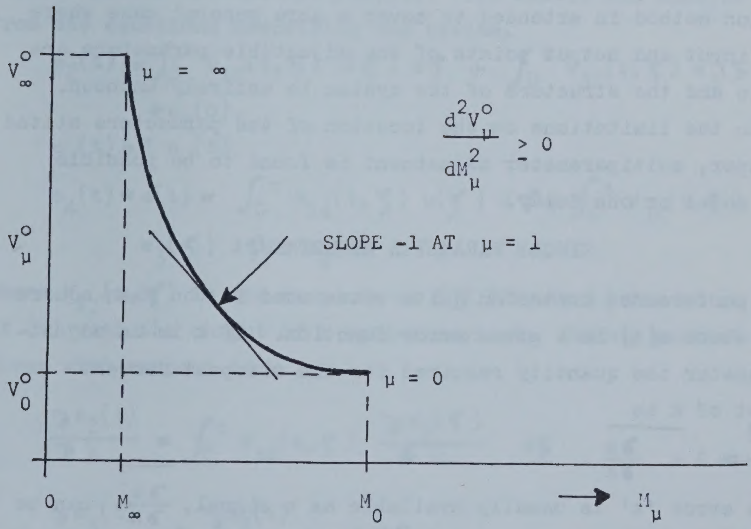


FIGURE 2

ADAPTIVE CONTROL USING TIME DELAY AND CORRELATION

KUMPATI S. NARENDRA
Dunham Laboratory
Yale University
New Haven, Connecticut

M. A. L. THATHACHAR
Indian Institute of Science
Bangalore, India

THEODORE S. BAKER
Dunham Laboratory
Yale University
New Haven, Connecticut

INTRODUCTION

A basic problem in self-optimizing or adaptive control systems using perturbation techniques is to get the components of the gradient of a performance criterion with respect to adjustable parameters as real time signals. A correlation technique has been employed for obtaining the gradient for specific configurations using a real model of the linear system[1] and has been modified to dispense with the model by the use of a delay.[2] In this paper the correlation method is extended to cover a more general case where only the input and output points of the adjustable parameters are accessible and the structure of the system is entirely unknown. Subject to the limitations on the location of the parameters stated in the paper, multiparameter adjustment is found to be possible with one model or one delay.

SINGLE PARAMETER ADJUSTMENT

The performance criterion quite often used is the mean square error $\overline{e^2}$ where e(t) is a given error function. If k is an adjustable parameter the quantity required for the steepest descent adjustment of k is

$$\frac{\partial \overline{e^2}}{\partial k} = 2\,\overline{e\,\frac{\partial e}{\partial k}}\,.$$

Since the error 'e' is usually available as a signal, $\frac{\partial \overline{e^2}}{\partial k}$ can be obtained by a correlation of e and $\frac{\partial e}{\partial k}$, if the latter is also generated as a real time signal. Even when a more general performance index $J = \overline{F(e, t)}$ is used, the crucial part in determining the gradient lies in obtaining $\frac{\partial e}{\partial k}$ as a signal.

Gradient in Parameter Space (Single Adjustable Parameter)

If a linear system (Figure 1) has an input u(t), error e(t), and adjustable parameter k, the partial derivative $\frac{\partial e(t)}{\partial k}$ may be obtained by using a model. If the model is at rest at time t_0 and the input to k, $e_2(t)$, is injected at a summing junction placed at the output of k in the model, every signal in the model (except the output at k) is the partial derivative of the corresponding signal in the system with respect to $k(t_0)$.

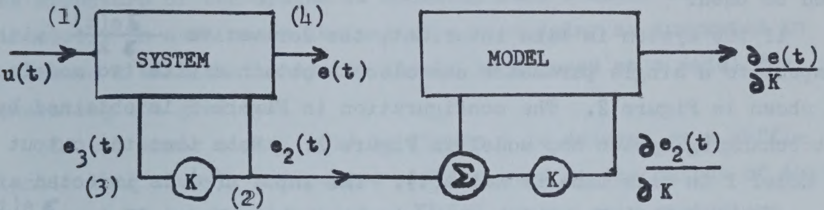

(1)
(4)
u(t)
SYSTEM
e(t)
MODEL
$\frac{\partial e(t)}{\partial K}$
$e_3(t)$
(3)
K
$e_2(t)$
(2)
Σ
K
$\frac{\partial e_2(t)}{\partial K}$

FIGURE 1

(Using the definition of a derivative, $\frac{\partial x(t)}{\partial k}$ can be found by evaluating the change in x(t) due to a small change in k. Since the system is linear and only the increment is required, the use of a model yields the desired result.) The result can also be obtained from the equations describing the system.

$$e_2(t) = \int_0^t W_{12}(t,\zeta)\, u(\zeta)\, d\zeta + \int_0^t W_{32}(t,\zeta)\, e_3(\zeta) d\zeta + e_2(0).$$

$$e_3(t) = K\, e_2(t) \tag{1}$$

$$e_4(t) = e(t) = \int_0^t W_{14}(t,\zeta)\, u(\zeta)\, d\zeta + \int_0^t W_{34}(t,\zeta)\, e_3(\zeta)\, d\zeta + e_4(0)$$

where $W_{ij}(t,\zeta)$ is the response at (j) (Figure 1) due to an impulse at (i) with the output of k open-circuited. Taking partial derivatives with respect to k

$$\frac{\partial e_2(t)}{\partial k} = \int_0^t W_{32}(t,\zeta)\, \frac{\partial e_3(\zeta)}{\partial k}\, d\zeta$$

$$\frac{\partial e_3(t)}{\partial k} = K\, \frac{\partial e_2(t)}{\partial k} + e_2(t) \tag{2}$$

$$\frac{\partial e(t)}{\partial k} = \int_0^t W_{34}(t,\zeta)\, \frac{\partial e_3(\zeta)}{\partial k}\, d\zeta$$

Therefore, the partial derivatives (even for time-varying systems) are seen to be the outputs of an exact model with zero input and $e_2(t)$ injected at the output of k.

MULTIPARAMETER ADJUSTMENT

When several parameters (n) have to be adjusted simultaneously, 'n' models are usually necessary. However, it can be shown that by a suitable choice of the location of the parameters fewer models need be used.

If the system is time invariant, the derivative $\frac{\partial e(t)}{\partial k}$ with respect to a single parameter can also be obtained with two models as shown in Figure 2. The configuration in Figure 2 is obtained by interchanging system and model in Figure 1. Note that the output of Model I in this case is not e(t). The input u(t) is injected at the output of k in Model I and the input to k in Model II is $\frac{\partial e(t)}{\partial k}$.

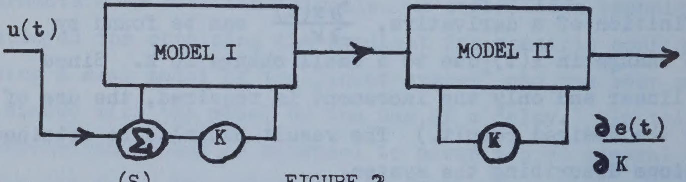


FIGURE 2

While this scheme using two models is not attractive for a single adjustable parameter, it has the advantage that it is applicable to systems with any number of gain parameters whose outputs end at a common summing point (S in Figure 2). If $k_i (i = 1, 2 \dots n)$ are the adjustable parameters, the elements of the gradient vector, $\frac{\partial e(t)}{\partial k_i}$ $(i = 1, \dots n)$ are the signals in Model II at the inputs of the corresponding gain parameters k_i.

A further simplification is possible when the outputs of the parameters are connected to the input as shown in Figure 3. In such a case, Model I of Figure 2 can be replaced by the system itself and hence one model is sufficient to determine all the components of the gradient.

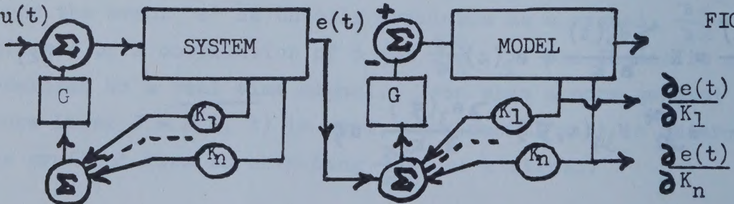

u(t)
Σ
SYSTEM
G
K_1
K_n
Σ
e(t)
+
-
Σ
MODEL
G
K_1
K_n
Σ
$\frac{\partial e(t)}{\partial K_1}$
$\frac{\partial e(t)}{\partial K_n}$

FIGURE 3

The determination of the gradient as indicated in Figure 3, using a single model, is also possible when the system varies slowly with time and does not change appreciably in an interval equal to the averaging time of the adaptive system.

USE OF SYSTEM AS ITS OWN MODEL

The results of the previous section involve the use of an exact model of the system. In a truly adaptive situation, where the structure of the system is unknown, such a model cannot be constructed. In such cases the use of a time delay as suggested in Reference 2 enables the system itself to be used as a model for generating the gradient.

If the error signal e(t) in Figure 3 is delayed by a sufficiently long time T, attenuated and injected at the output of the k_i's as in Figure 4, the inputs $\eta_i(t)$ to the gain parameters would contain two components which can be made to be uncorrelated. The larger components may be considered to be due to the input u(t), while the smaller components containing the information regarding the gradient $\frac{\partial e(t-T)}{\partial k_i}$ may be considered to be due to the delayed and attenuated error signal.

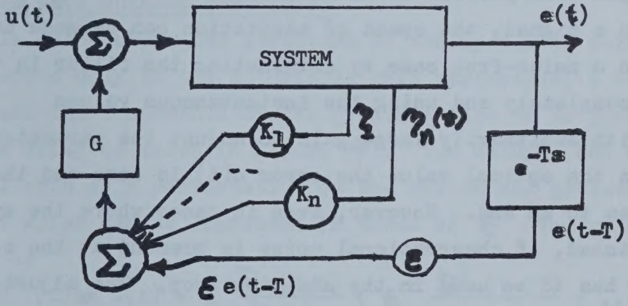


FIGURE 4

Hence the gradient of the performance index in parameter space is obtained by correlating the signals $\eta_i(t)$ with e(t-T). The error signal fed back into the system is kept sufficiently small (i.e., $\epsilon \ll 1$) so that it does not significantly distort the performance index or its derivatives.

ADAPTIVE SYSTEMS

In the previous sections the partial derivatives $\frac{\partial e}{\partial k_i}$ were generated as signals using a model of the system or by a time delay in the feedback path. Once $\frac{\partial e}{\partial k_i}$ are available as signals, the correlation $\overline{e \frac{\partial e}{\partial k_i}}$ can be used to adjust k_i down the gradient of $\overline{e^2}$ in parameter space so that $\overline{e^2}$ has a minimum value. In practice the "correlator" is a filter which integrates or takes a time average of the product $e \frac{\partial e}{\partial k}$ depending on whether the adaptive loop is to be a servo or a regulator. It also establishes the gain of the feedback loop and rejects signal components which would interfere with the adaptive process when such components exist. The rate of adaptation is decided by the bandwidth of the adaptive signal which generally contains mainly low frequency components.

Systems with Known Structures

In identification problems where an unknown system has to be identified in terms of a specific mathematical model (or white box) the above method can be used to generate the adaptive signal. Two identical models are used, one to generate the error e(t) and the other, the partial derivatives $\frac{\partial e(t)}{\partial k_i}$. If the mathematical model chosen is such that it can match the given system exactly (noise-free case) the performance criterion $\overline{e^2(t)}$ is zero and is independent of the input when perfect match is achieved. Since $\frac{\partial e^2(t)}{\partial k_i}$ is available as a signal, the speed of adaptation can be made arbitrarily high in a noise-free case by eliminating the filter in the adaptive loop completely and using the instantaneous values $e(t) \frac{\partial e(t)}{\partial k_i}$ with arbitrarily large gain to adjust the parameter k_i. When k_i reaches the optimal value the error e(t) is zero and the adaptation comes to an end. However, even in cases where the systems can be matched, if observational noise is present at the outputs, a filter has to be used in the adaptive loop. The adjustment of the parameters of a third order system with additive Gaussian noise at the output to match a given third order system is shown in Figure 7.

Situations where the model cannot match the system exactly are of much greater interest and the optimum values of the parameters are intimately related to the statistical parameters describing the input. The components of the gradient in parameter space cannot be determined, in such a case, in less time than is required to identify the parameters (or statistics) of the input signal. For random inputs this sets an upper bound on the speed of adaptation.

A special relevant example of the technique is given by the differential equation problem[3] shown in Figure 5 where the coefficients of the differential equation have to be adjusted to minimize a performance criterion $\overline{e^2}$ (or $\frac{1}{T}\int_0^T e^2(t)dt$ for deterministic inputs).

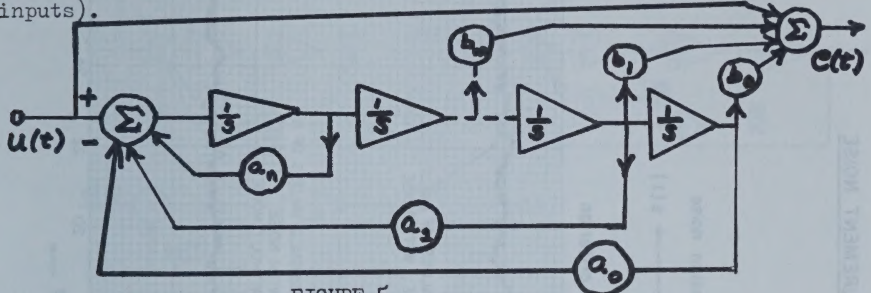


FIGURE 5

The coefficients b_i ($i = 1, \ldots m$) as also the coefficients a_i ($i = 1, \ldots n$) are seen to end at common summing points and hence the correlation technique can be used to adjust the various parameters simultaneously using a single model. The adjustment of the parameters of a third order system in parameter space is shown in Figure 8.

Systems with Unknown Structures

In cases where the structure of the system is unknown and one has access only to the adjustable gain parameters, a model cannot be used and the components of the gradient have to be generated using a time delay as shown in Figure 4. The delayed and attenuated signal acts as a perturbation on the system and produces an error $e^*(t)$ which can be expressed in terms of ϵ (the gain in the delay loop) as

$$e^*(t) = e(t) + \epsilon \frac{\partial e(t-T)}{\partial k} + \epsilon^2 \frac{\partial^2 e(t-2T)}{\partial k^2} + \ldots \qquad (3)$$

[This results from expanding the transfer function $W(s, k + \epsilon e^{-sT})$ around the unperturbed transfer function.]

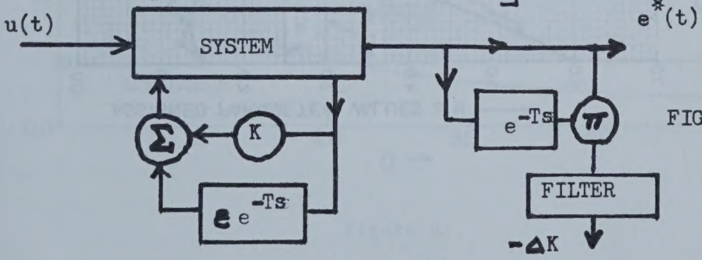


FIGURE 6

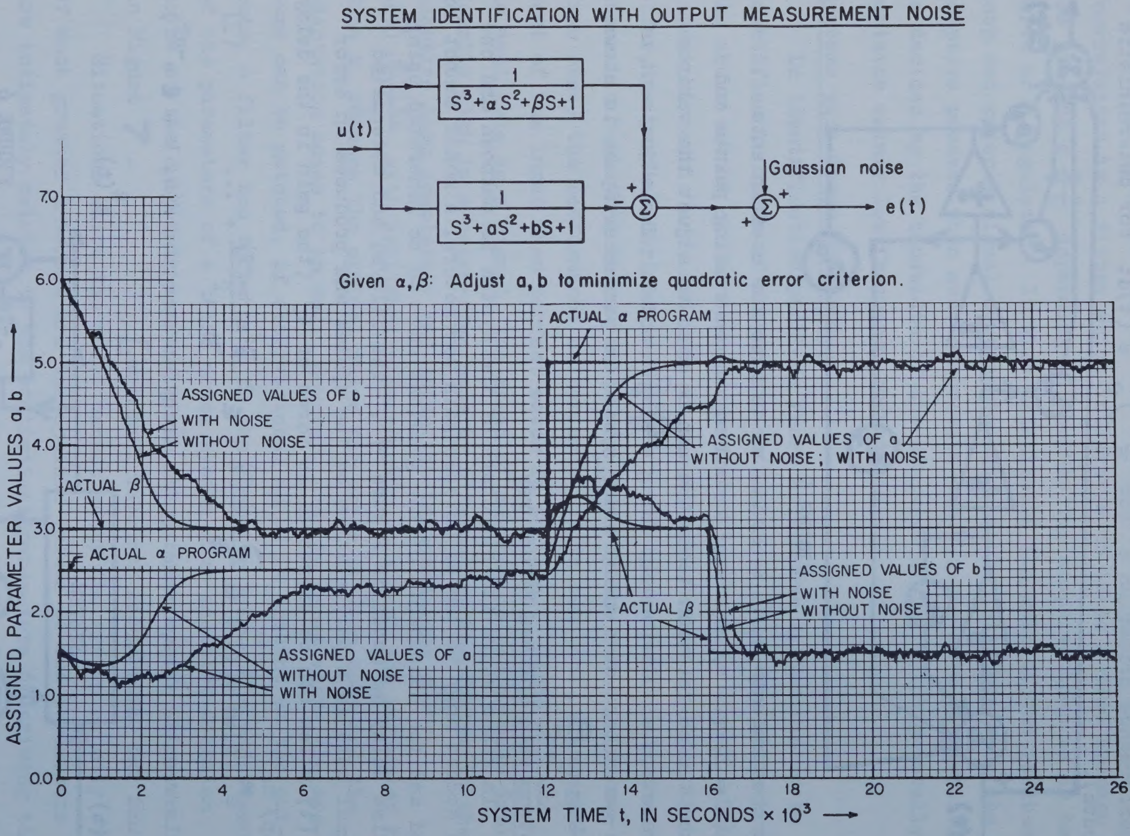
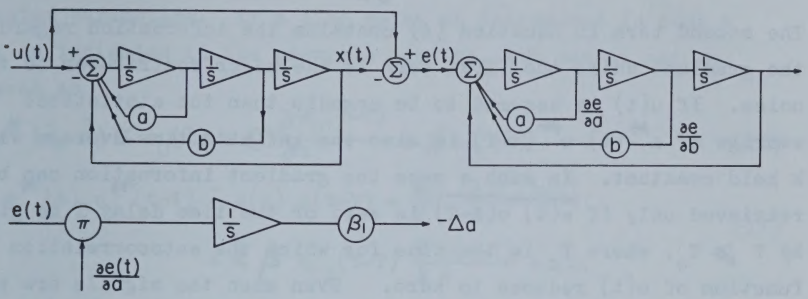

SYSTEM IDENTIFICATION WITH OUTPUT MEASUREMENT NOISE
u(t)
$\frac{1}{S^3+\alpha S^2+\beta S+1}$
$\frac{1}{S^3+aS^2+bS+1}$
+
−
Σ
Gaussian noise
+
+
Σ
e(t)
Given α, β: Adjust a, b to minimize quadratic error criterion.
ASSIGNED PARAMETER VALUES a, b
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
SYSTEM TIME t, IN SECONDS × 10^3
ASSIGNED VALUES OF b
WITH NOISE
WITHOUT NOISE
ACTUAL β
ACTUAL α PROGRAM
ASSIGNED VALUES OF a
WITHOUT NOISE
WITH NOISE
ACTUAL α PROGRAM
ASSIGNED VALUES OF a
WITHOUT NOISE; WITH NOISE
ACTUAL β
ASSIGNED VALUES OF b
WITH NOISE
WITHOUT NOISE

Figure 7

ADJUSTMENT OF PARAMETERS OF A THIRD ORDER SYSTEM USING A MODEL.


u(t)
+
−
Σ
1/s
1/s
1/s
x(t)
−
Σ
+ e(t)
Σ
1/s
1/s
1/s
a
b
a
b
∂e/∂a
∂e/∂b
e(t)
π
1/s
β1
−Δa
∂e(t)/∂a

Square Wave Input

Quadratic Error Criterion

$\dddot{x} + a\ddot{x} + b\dot{x} + x = u(t)$

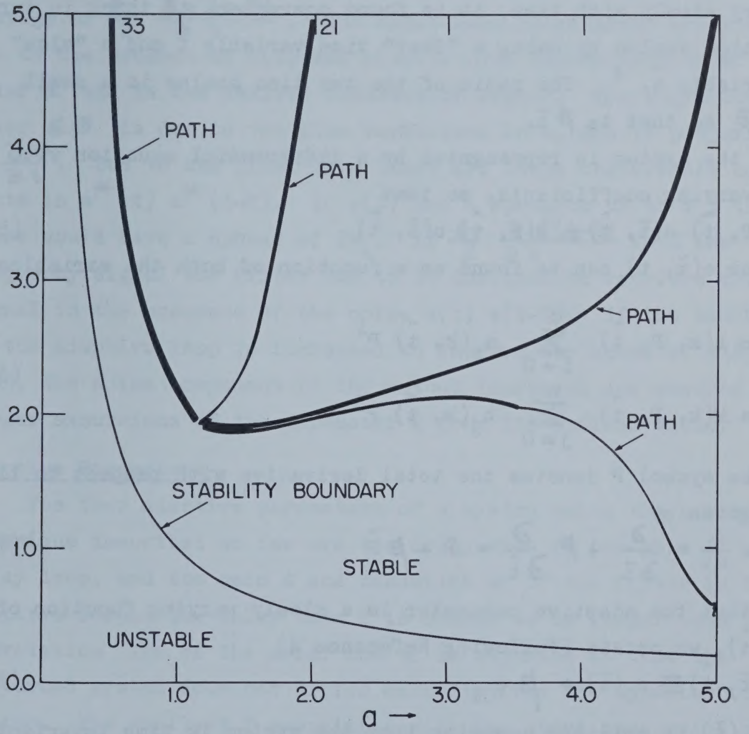


Figure 8

If $e^{*}(t)$ is delayed by time T and multiplied by $e^{*}(t)$, one obtains

$$e^{*}(t)\, e^{*}(t-T) \cong e(t)\, e(t-T) + \frac{\epsilon}{2}\left[\frac{\partial}{\partial k}\, e^{2}(t-T)\right] + \epsilon\, e(t)\, \frac{\partial}{\partial k}\, e(t-2T) + \ldots \qquad (4)$$

The second term in Equation (4) contains the information regarding the gradient while the first term is the major contributor to the noise. If $u(t)$ is assumed to be ergodic then the statistical average of $e^{*}(t)\, e^{*}(t-T)$ is also the infinite time average with k held constant. In such a case the gradient information can be retrieved only if $\overline{e(t)\, e(t-T)}$ is zero or the time delay T is given by $T \geq T_c$, where T_c is the time for which the autocorrelation function of $e(t)$ reduces to zero. Even when the signals are passed through a filter, it is found that this condition is desirable for most systems (refer to the next section).

Time Varying Systems. When the system (or input) characteristics vary slowly with time, it is found convenient to think in terms of two time scales by using a "fast" time variable $\tilde{t}$ and a "slow" time variable $\hat{t}$.[4] The ratio of the two time scales is a small number β so that $\hat{t} = \beta\, \tilde{t}$.

If the system is represented by a differential equation with slowly varying coefficients, so that

$$L(P, \hat{t})\, e(\tilde{t}, \hat{t}) = M(P, \hat{t})\, u(\tilde{t}, \hat{t}) \qquad (5)$$

the error $e(\tilde{t}, \hat{t})$ can be found as a function of both the variables $\tilde{t}$ and $\hat{t}$.

$$\left.\begin{aligned} L &= L(k, P, \hat{t}) \quad \sum_{i=0}^{n} a_i(k, \hat{t})\, P^{i} \\ M &= M(k, P, \hat{t}) \quad \sum_{j=0}^{n} b_j(k, \hat{t})\, P^{j} \end{aligned}\right\} \qquad (6)$$

where the symbol P denotes the total derivative with respect to time and is given by

$$P = \frac{d}{dt} = \frac{\partial}{\partial \tilde{t}} + \beta \frac{\partial}{\partial \hat{t}} = \tilde{P} + \beta \hat{P} \qquad (7)$$

Noting that the adaptive parameter is a slowly varying function of time $k(\hat{t})$, we obtain (following Reference 4)

$$e(\tilde{t}, \hat{t}) \cong e_0(\tilde{t}) + \beta\, e_{01} \qquad (8)$$

where $e_0(\tilde{t})$ is obtained assuming that the system is time invariant ($\hat{t}$ is held constant) and e_{01} is given by

$$e_{01} = -\frac{\partial k(\hat{t})}{\partial \hat{t}} \left[\frac{\partial L/\partial \tilde{P}}{L(\tilde{P})}\right] \frac{\partial e_0(\tilde{t})}{\partial k} \qquad (9)$$

e_{01} represents the component of the error due to slow variations in the system parameters. If a time delay is introduced in such a system as indicated in the previous section, the output may be expressed as

$$e^*(\tilde{t}, \hat{t}) \cong e_0(\tilde{t}) + \epsilon \frac{\partial e_0(\tilde{t}-T)}{\partial k} + \beta e_{01}(\tilde{t}) \qquad (10)$$

Hence $e^*(t)\, e^*(t-T) \simeq e(t)\, e(t-T) + \frac{\epsilon}{2} \frac{\partial e^2(t-T)}{\partial k}$

$$+ \epsilon \beta\, e_{01}(t-T) \frac{\partial e(t-T)}{\partial k} + \dots \qquad (11)$$

(where t denotes the fast time variable.) Since $k(\hat{t})$ is not fixed, infinite time averages are not possible. A finite time average is provided by the filter in the adaptive loop and it is required that the output of the filter be a good measurement of the statistical average or $\frac{\epsilon}{2} \frac{\partial e^2(t-T)}{\partial k}$. The first term is of order unity and is due to the product of e(t) and e(t-T). The second term is of order ϵ and is the desired information signal. The third term of order $\epsilon\beta$ is due to the slow variations in k, and is proportional to $\frac{\partial k}{\partial t}$. Due to the first term there are large undesirable components in $e^*(t)\, e^*(t-T)$. If e(t) has a spectrum up to W_e, this noise would have a spread of $2W_e$. If $\frac{\epsilon}{2} \frac{\partial e^2(t-T)}{\partial k}$ is a low frequency signal the filter has to be designed to retrieve this signal in the presence of the noise e(t) e(t-T). If the bandwidth of the adaptive loop is increased to improve the speed of adaptation, the noise component in the output increases and results in larger excursions of the parameter k from its optimal value.

Adaptive Parameters

The four adaptive parameters of a system using the perturbation technique described so far are the delay time T, the gain ϵ of the delay loop, and the gain A and bandwidth α of the filter in the adaptive loop. The delay time T is chosen to be larger than the correlation time of the error and ϵ is adjusted so that the perturbed system does not differ markedly from the unperturbed system. The smallest T compatible with the above constraint should be chosen since a large value of T implies a considerable lag in the adaptive process. If an integrator $\frac{A}{s}$ is used in the feedback path, both gain and bandwidth of the adaptive loop are controlled by the parameter A. A is chosen from a priori knowledge of the variations in the system and the error spectrum to obtain the optimal signal k(t). The adaptive control of a third order system starting from different initial conditions in parameter space is shown in Figure 10.

Perturbation Methods

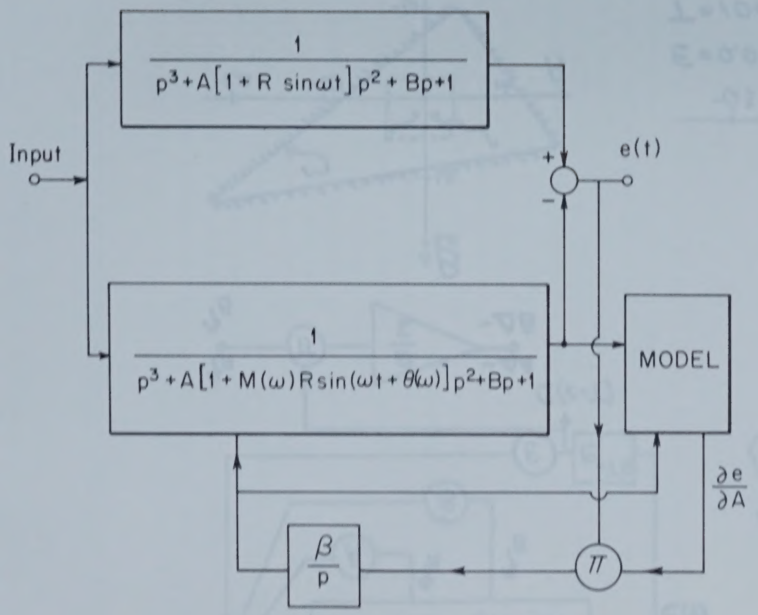
The method proposed in this paper is seen to have many features in common with the well-known parameter perturbation scheme[5] for adaptive systems. The authors feel that the two schemes are mathematically much more similar than is readily apparent, and investigation in this area is being continued.

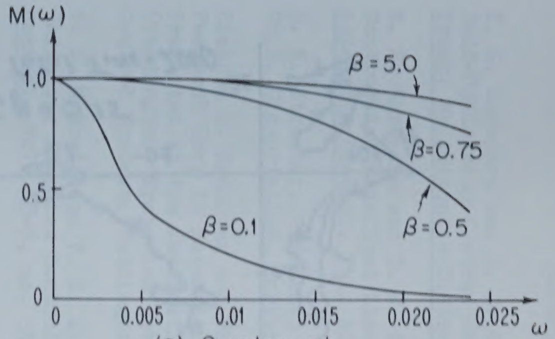
While parameter perturbation methods work for nonlinear systems, the proposed method is restricted to linear systems and gain type parameters at the present time. However, through the utilization of favorable system topology one model or one delay can be made to serve for the adjustment of different parameters. A comparable scheme is not available in parameter perturbation schemes at the present time.

CONCLUSION

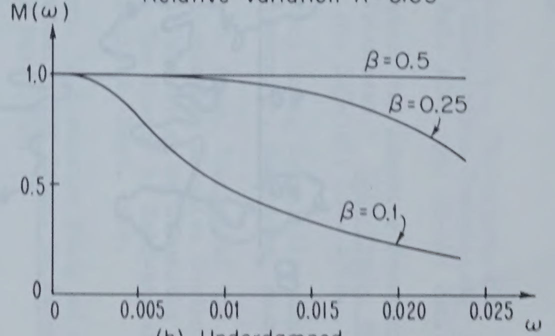
A new method has been presented in this paper for adjusting the parameters of a system to optimize a performance criterion. When a model of the system can be constructed, the partial derivatives of the error may be generated explicitly as real time signals and used to adjust the control parameters along the gradient of the performance index in parameter space. When the structure of the system is unknown the scheme uses a time delay and the functions $\frac{\partial e(t)}{\partial k_i}$ are generated as uncorrelated signals within the system. The determination of the gradient in parameter space reduces to a signal detection problem and involves a trade-off between speed of adaptation and the deviation from the minimum error.

The scheme proposed strongly resembles adaptive methods using parameter perturbation. While parameter perturbation can be used for nonlinear systems, the method described above applies only to linear systems at the present time. Its main advantage, however, lies in the fact that several parameters can be simultaneously adjusted using a single time delay.


Input
1
p³+A[1+R sinωt]p²+Bp+1
+
−
e(t)
1
p³+A[1+M(ω)Rsin(ωt+θ(ω))]p²+Bp+1
MODEL
∂e/∂A
π
β/p

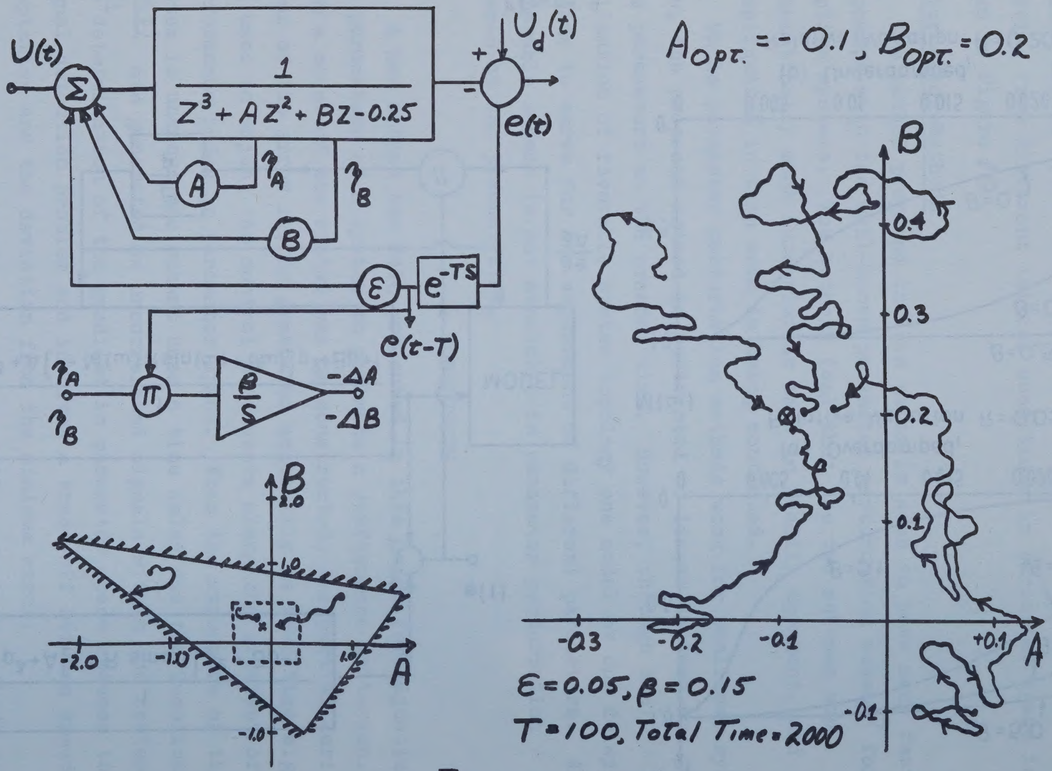


(a) Overdamped;
Relative Variation R=0.05



(b) Underdamped,
Relative Variation R=0.20

Figure 9


U(t)
Σ
$\frac{1}{Z^3 + AZ^2 + BZ - 0.25}$
$U_d(t)$
+
-
e(t)
η_A
η_B
A
B
ε
e^{-TS}
e(t-T)
η_A
η_B
π
$\frac{\beta}{S}$
-ΔA
-ΔB
B
2.0
1.0
-2.0
-1.0
1.0
-1.0
A
B
0.4
0.3
0.2
0.1
-0.1
-0.3
-0.2
-0.1
+0.1
A

$A_{Opt.} = -0.1, B_{Opt.} = 0.2$

$\varepsilon = 0.05, \beta = 0.15$

$T = 100$, Total Time = 2000

Figure 10

REFERENCES

1. K. S. Narendra and L. E. McBride, Jr., "Multiparameter Self-Optimizing Systems Using Correlation Techniques." IEEE Trans. on Automatic Control, Vol. AC-9, pp. 31-38; Jan. 1964.

2. K. S. Narendra and D. N. Streeter, "An Adaptive Procedure for Controlling Undefined Linear Processes." IEEE Trans. on Automatic Control, Vol. AC-9, pp. 545-548; Oct. 1964.

3. D. N. Streeter, "Adaptive Control Using Correlation Methods." Ph.D. Dissertation, Harvard University, Cambridge, Mass. 1964.

4. Richard E. Kronauer and Philip G. Drew, "Design of the Adaptive Feedback Loop in Parameter-Perturbation Adaptive Controls." Presented at the International Federation on Automatic Control (Teddington) Symposium 1965 on The Theory of Self-Adaptive Control Systems. Teddington, England, Sept. 1965.

5. R. J. McGrath and V. C. Rideout, "A Simulator Study of a Two-Parameter Adaptive Control System." IRE Trans. on Automatic Control, Vol. AC-6, pp. 35-42; Feb. 1961.

ANALYSIS OF DISTRIBUTED PARAMETER NETWORKS - A GENERAL METHOD

DANIEL DICKER
State University of New York at Stony Brook
Stony Brook, New York

ABSTRACT

A general method is presented for obtaining approximate analytical solutions for distributed parameter networks where the parameters may be functions of the spatial variable. By means of this technique, an explicit approximate solution for the output of a rather general network is presented. For those cases where exact separable solutions exist the method yields the correct solutions.

INTRODUCTION

The distributed network considered in this paper is a fairly general one, and characterized by the "transmission line" equations [1] [2]

$$\begin{aligned} -\partial e(x,t)/\partial x &= ri(x,t) + \ell\, \partial i(x,t)/\partial t \\ -\partial i(x,t)/\partial x &= ge(x,t) + c\, \partial e(x,t)/\partial t \end{aligned} \qquad (1)$$

For certain limiting cases general solutions of (1) are available, e.g. high-frequency lines and submarine cables, where the governing partial differential equations reduce to one of the standard types of applied mathematics - the wave equation or the heat equation. Examples of such solutions may be found in the references cited. For the full equations (1) and where the termination is neither open circuited nor short circuited, the boundary value problem is not separable, hence an analytical solution is generally not available. Furthermore, in the event the line parameters r, ℓ, g, c are nonconstant the difficulty is compounded so that solutions for specific problems must be obtained numerically either by analog or digital computer.

A general method has been developed for obtaining approximate analytical solutions for (1) where the parameters may be dependent upon x but independent of time. By manipulating the differential equation it will be shown that the boundary-value problem can be reduced to a Sturm-Liouville system. The latter is then solved by means of the Bubnov-Galerkin method to yield the current and voltage as functions of x and t in terms of an infinite series, the coefficients of which are obtainable from a system of linear algebraic equations. By truncating the algebraic system approximations to any degree of precision may then be obtained. For those cases where exact solutions exist the algebraic system is easily solved in closed form.

REDUCTION OF BOUNDARY-VALUE PROBLEM TO STURM-LIOUVILLE SYSTEM

Consider the first equation of (1). Viewing it as a first-order nonhomogeneous differential equation on the current with time as the independent variable it is easily solved by standard techniques to yield

$$i(x,t) = \exp(-rt/\ell) \int_0^t \exp(r\tau/\ell)\partial e(x,\tau)/\partial x \, d\tau + C_1 \exp(-rt/\ell) \qquad (2)$$

The coefficient C_1 may be a function of x. Although it is not necessary to assume initial quiescent conditions for this method, as a convenience in this study it will be assumed that e(x,0) and i(x,0) are zero. Under this assumption C_1 vanishes and the formal solution for the current becomes

$$i(x,t) = \exp\big(-r(x)t/\ell(x)\big) \int_0^t \exp\big(r(x)\tau/\ell(x)\big)\partial e(x,\tau)/\partial x d\tau \qquad (3)$$

Next substituting for the current from (3) into the second equation of (1) yields

$$-\partial/\partial x[\exp\big(-r(x)t/\ell(x)\big) \int_0^t \exp\big(r(x)\tau/\ell(x)\big)\partial e(x,\tau)/\partial x d\tau] = g(x)e(x,t) + c(x)\partial e(x,t)\partial t \qquad (4)$$

which is an integro-partial-differential equation on the variable e(x,t). Applying next the Laplace Transformation to (4) gives

$$\frac{d}{dx}\,\frac{dE(x;s)/dx}{s\ell(x) + r(x)} = [g(x) + sc(x)]\, E\,(x;s) \qquad (5)$$

where $\mathcal{L}[e(x,t)] \equiv \int_0^\infty \exp(-st)e(x,t)\,dt \equiv E(x;s)$ [3]

The problem has thus been reduced to the solution of the self-adjoint differential equation (5). To prove that the original boundary value problem is equivalent to a Sturm-Liouville system requires a brief discussion of the boundary conditions, as follows.

Reduction of Boundary Conditions

Consider the terminal end of a transmission line consisting of R, L and C elements. The line voltage and current at the end of the line, x = b, are designated e(b,t) and i(b,t) respectively. Consider next an arbitrary path across the line in the termination circuit. It is easily shown that the following relationship is satisfied

$$\dot{e}(b,t) = \sum_{k=1}^{n} [I_{jk}/C_k + R_k\dot{I}_{jk} + L_k\ddot{I}_{jk}] \qquad (6)$$

where I_{jk} is the current in the k^{th} element of j^{th} path; and j=1,2,...m, where m is the number of such paths. Hence there are "m" equations (6). Furthermore, summing currents at each junction supplies the requisite additional number of equations required to reduce the system to a single differential equation involving e(b,t) and i(b,t) only, viz.,

$$(\alpha_{n+1}d^{n+1}/dt^{n+1}+\alpha_n d^n/dt^n+\ldots.\alpha_o)i(b,t) = (\beta_n d^n/dt^n+\beta_{n-1}d^{n-1}/+\ldots\beta_o)e(b,t) \qquad (7)$$

Combining (7) above with the first equation of (1) evaluated at x = b gives

$$-(\alpha_{n+1}D^{n+1}+\alpha_n D^n+..\alpha_o)\partial e(b,t)/\partial x = (r(b)+\ell(b)D)\ (\beta_n D^n+....\beta_o)e(b_1 t) \qquad (8)$$

where $D \equiv d/dt$. Finally, applying the Laplace Transformation to (8) gives for the appropriate boundary condition at $x = b$

$$-h(s)dE(b;s)/dx = E(b;s) \qquad (9)$$

where $$h(s) = [\alpha_{n+1}s^{n+1}+\alpha_n s^n+..\alpha_o]/[\ell(b)\beta_n s^{n+1}+(r(b)\beta_n+\ell(b)\beta_{n-1})s^n +....r(b)\beta_o] \qquad (10)$$

At the sending end the most usual boundary condition is the specification of the voltage,

$$e(0,t) = f(t) \qquad (11)$$

the Laplace transformation of which is

$$E(0;s) = \bar{f}(s) \qquad (12)$$

The boundary-value problem consisting of (5), (9) and (12) is a Sturm-Liouville problem* [4] for which it is known that not only does a unique solution exist, but also that there exists a full set of discrete eigenfunctions, so that the solution may be expressed in terms of an expansion of these eigenfunctions. However, there are relatively few solutions available in terms of tabulated functions, and these correspond to very special functions for the line parameters.** In the following section the Bubnov-Galerkin technique will be described, which will be employed to generate a convergent approximate eigenfunction expansion with arbitrary line parameters.

---

* The boundary-value problem is also a Sturm-Liouville problem when the boundary condition at $x = 0$ is similar to (9), i.e. $h'(s)dE(0;s)/dx = E(0;s)$, and, of course, when the boundary condition at $x = b$ is similar to (12), i.e. $E(b;s) = \bar{f}'(s)$.

** One nontrivial example would be, for instance, $\ell(x)$ = const., $r(x)$ = const., $g(x)$ = const., $c(x) = \cos 2x$, the resulting solution for $E(x;s)$ being an expansion in Mathieu functions.

SOLUTION BY THE BUBNOV-GALERKIN METHOD

Bubnov-Galerkin Method

The following is a brief outline of the technique; for a detailed mathematical presentation the reader is referred to Mikhlin [5], and numerous examples in the study of differential equations and continuum mechanics may be found in Kantorovich and Krylov [6]. Consider a differential equation

$$Au = q \tag{13}$$

where A is a differential operator, u the dependent variable and q a nonhomogeneous term. The approximate solution is assumed to be of the form

$$u^{(n)} = \Sigma_{k=1}^{n} \rho_k \varphi_k \tag{14}$$

where the φ_k are elements of a complete set and ρ_k are coefficients. It is necessary that $u^{(n)}$ satisfy the prescribed boundary conditions. The Bubnov-Galerkin method consists in equating to zero the scalar products

$$\int_\tau (Au^{(n)} - q)\, \varphi_k \, d\tau \;, \quad k = 1, 2 \ldots n$$

where τ is the domain of definition of u, and solving the resulting linear algebraic simultaneous system for the coefficients, ρ_k.

Solution of the Network Problem

For the problem at hand, (5), (9), (12), it is convenient to introduce a new variable, u, such that

$$E = \lambda + u \tag{15}$$

where λ is a term which satisfies the given boundary conditions; u, therefore, satisfies homogeneous boundary conditions. After substituting (15) into (5), (9), (12), and in (5) carrying out the differentiations and multiplying by $(\ell s + r)^2$ there results as the problem on u

$$Au = -A\lambda \tag{16}$$
$$h\,du/dx + u = 0 \qquad \text{at } x = b \tag{17}$$
$$u = 0 \qquad \text{at } x = 0 \tag{18}$$

where

$$A \equiv (\ell s+r)\, d^2/dx^2 - (\ell' s+r')\, d/dx - (\ell s+r)^2(g+cs) \tag{19}$$

A simple choice for λ is

$$\lambda = \bar{f}\,[1 - x/(h - b)] \tag{20}$$

and a suitable collection of functions which are elements of a set complete in (0,b) is

$$\varphi_1 = x + (h - b)x^2/b(2h - b) \quad , \quad \varphi_n = x(x - b)^n; \; n > 1 \tag{21}$$

It is easily verified that (20) satisfies (9) and (12) and that each φ_k of (21) satisfies (17) and (18).

As is customary, define the scalar product of two functions, v(x), w(x) as

$$(v,w) \equiv \int_0^b vw\,dx \tag{22}$$

Similarly the energy product is

$$[v,w] \equiv \int_0^b w\,Av\,dx \tag{23}$$

where A is an operator. In terms of these definitions the simultaneous system of algebraic equations is given simply by

$$[u^{(n)} - \lambda,\ \varphi_k] = 0 \qquad k = 1,\ 2 \ldots n \tag{24}$$

where

$$u^{(n)} = \rho_1\varphi_1 + \rho_2\varphi_2 + \ldots \rho_n\varphi_n \tag{25}$$

The approximate solution is obtained by selecting a value of n, solving for the coefficients ρ_k, and inverting $u^{(n)} + \lambda$ to obtain e(x,t).

Of great importance in applications is the output, e(b,t).* An examination of the assumed solution for E(x;s), (15), (25), (21), shows that the series terminates after the first term and that

$$E(b;s) = \lambda + \rho_1\varphi_1(b;s) \tag{26}$$

which yields, for the output,

$$e(b,t) = \mathcal{L}^{-1}[\lambda + \rho_1\varphi_1] \tag{27}$$

This last result is of importance as it is an explicit formula for the output, for a fairly arbitrary network, in terms of known functions λ and φ_1 as given by (20) and (21). Because of the nature of the termination circuit h(s) is simply the ratio of two polynomials in s, the degree of each of which depends upon the complexity of the termination. Likewise, the coefficient ρ_1 is a ratio of polynomials in s. In general, the only nonpolynomial function of s which appears in (26) is the input $\overline{f}(s)$. Therefore the inverse Laplace transform is not difficult to obtain once the polynomials have been factored.

It is to be noted, however, that the coefficients, ρ_k, are so called "flexible" coefficients. The value of ρ_1 obtained from a one term approximation will differ somewhat from ρ_1 obtained from an 'n' term approximation. Designate this latter coefficient $\rho_1^{(n)}$. It is known [5] that $\rho_1^{(n)}$ converges, i.e.

$$\rho_1^{(n)} \rightarrow \rho_1, \text{ and } \rho_k^{(n)} \rightarrow \rho_k \tag{28}$$

---

* When it is rather the current i(b,t) that is desired, equation (3) together with the solution for e(b,t) will provide the required result. Alternatively, the entire problem may be formulated in terms of i(x,t).

The power of the Bubnov-Galerkin algorithm lies not only in (28) but also in that the solutions so constructed generally yield very close approximations to the exact solution, even when the number of terms included in the approximate solution (25) is small [7], [8].

Various approximations will be exhibited for the solution to (5), (9), (12). For a one term approximation

$$u^{(1)} = \rho_1^{(1)}\varphi_1 \tag{29}$$

where φ_1 is given by (21). Substituting in (24) gives

$$[u^{(1)}, \varphi_1] = [\lambda, \varphi_1]$$

and utilizing (29) there results

$$\rho_1^{(1)} = [\lambda, \varphi_1] \,/\, \|\varphi_1\|^2$$

where $\|\varphi_1\|$ is the so-called "energy norm" defined by $\|\varphi_k\| \equiv [\varphi_k, \varphi_k]$. Similarly the two term approximation, given by

$$u^{(2)} = \rho_1^{(2)}\varphi_1 + \rho_2^{(2)}\varphi_2$$

together with (24) gives the simultaneous system

$$[u^{(2)}, \varphi_1] = [\lambda, \varphi_1]$$
$$[u^{(2)}, \varphi_2] = [\lambda, \varphi_2]$$

from which it follows directly that

$$\rho_1^{(2)} = \frac{[\lambda,\varphi_1]\|\varphi_2\|^2 - [\lambda,\varphi_2][\varphi_2,\varphi_1]}{\|\varphi_1\|^2\|\varphi_2\|^2 - [\varphi_1,\varphi_2][\varphi_2,\varphi_1]}, \quad \rho_2^{(2)} = \frac{\|\varphi_1\|^2[\lambda,\varphi_2] - [\varphi_1,\varphi_2][\lambda,\varphi_1]}{\|\varphi_1\|^2\|\varphi_2\|^2 - [\varphi_1,\varphi_2][\varphi_2,\varphi_1]}$$

In an entirely analagous manner the coefficients for the nth approximation

$$u^{(n)} = \rho_1^{(n)}\varphi_1 + \rho_2^{(n)}\varphi_2 + \dots \rho_k^{(n)}\varphi_k + \dots \rho_n^{(n)}\varphi_n$$

are given by

$$\rho_k^{(n)} = \begin{vmatrix} \|\varphi_1\|^2 & [\varphi_2\varphi_1] & \cdots & [\varphi_{k-1},\varphi_1] & [\lambda,\varphi_1] & [\varphi_{k+1},\varphi_1] & \cdots & [\varphi_n,\varphi_1] \\ [\varphi_1,\varphi_2] & \|\varphi_2\|^2 & \cdots & [\varphi_{k-1},\varphi_2] & [\lambda,\varphi_2] & [\varphi_{k-1},\varphi_2] & \cdots & [\varphi_n,\varphi_2] \\ [\varphi_1,\varphi_n] & [\varphi_2,\varphi_n] & \cdots & [\varphi_{k-1},\varphi_n] & [\lambda\varphi_n] & [\varphi_{k+1},\varphi_n] & \cdots & \|\varphi_n\|^2 \end{vmatrix}$$
$$\div \begin{vmatrix} \|\varphi_1\|^2 & [\varphi_2,\varphi_1] & \cdots\cdots & [\varphi_n,\varphi_1] \\ [\varphi_1,\varphi_2] & \|\varphi_2\|^2 & \cdots\cdots & [\varphi_n,\varphi_2] \\ \vdots & \vdots & & \vdots \\ [\varphi_1,\varphi_n] & [\varphi_2,\varphi_n & \cdots\cdots & \|\varphi_n\|^2 \end{vmatrix} \tag{30}$$

It is clear that when it is possible to choose a complete set which is orthogonal in energy, (30) degenerates into a noncoupled system and the coefficients do not change from one approximation to the next, i.e.

$$\rho_k^{(n)} = \rho_k = [\lambda, \varphi_k] \,/\, \|\varphi_k\|^2 \; *$$

The problems for which this is possible are the "separable" ones; however, these are the exceptions rather than the rule.

Finally, the solution may be written

$$e(x,t) = \mathcal{L}^{-1}[\lambda] + \Sigma_{k=1}^{n} \mathcal{L}^{-1}[\rho_k^{(n)} \varphi_k]$$

---

* These are, of course, the Fourier coefficients.

REFERENCES

1. McLachlan, N.W.: Complex Variable Theory and Transform Calculus. Cambridge University Press, Second Edition, p. 235.

2. Miller, Kenneth S.: Partial Differential Equations. Prentice-Hall, Inc., 1953, p. 24.

3. Churchill, R.V.: Modern Operational Mathematics in Engineering. McGraw-Hill Book Company, Inc., 1944.

4. Sagan, H.: Boundary and Eigenvalue Problems in Mathematical Physics. John Wiley & Sons, Inc., 1961, Chapter V.

5. Mikhlin, S. G.: Variational Methods in Mathematical Physics. The Macmillan Company, 1964, p. 448.

6. Kantorovich, L V. and Krylov, V.I.: Approximate Methods of Higher Analysis. P. Noordhoff, Gromingen (1958).

7. Dicker, D. and Friedman, M. B.: Solutions of Heat-Conduction Problems with Nonseparable Domains. Jour. App. Mech., Vol. 30, series E, No. 4 (Dec.) 1963, p. 493.

8. Dicker, D. and Friedman, M. B.: Heat Conduction in Elliptical Cylinders and Cylindrical Shells. AIAA Jour., Vol. 1, No. 5 (May) 1963, p. 1139.

FREQUENCY CHARACTERISTICS OF A DOUBLE-KELVIN TRANSMISSION LINE NETWORK*

B.A.SHENOI
Department of Electrical Engineering
University of Minnesota
Minneapolis,Minnesota

ABSTRACT

The transfer functions of uniform and nonuniform distributed RC networks that have been studied extensively under various terminal conditions are found to exhibit either low-pass or high-pass characteristics only. Notch filters have been realized by connecting a resistance in series with a three-terminal low-pass distributed RC network. The well-known scheme for obtaining a band-pass filter characteristic is that of a high-gain amplifier with such a notch filter in the feedback path. This paper reports briefly on a study of the band-pass characteristics realizable by a four-layer distributed RC network without a series resistance and the amplifier scheme mentioned above. Such a passive two-port network is considered to be a better model for monolithic integrated circuits which are fabricated by a double diffusion process and are suitable for a cascaded network structure. Sufficient data has been obtained to show how the frequency characteristics can be varied by controlling the ratio of resistances and capacitances of the network.

It is also shown that null frequency characteristics with high dc attenuation and oscillator characteristics can be obtained from this network.

INTRODUCTION

Several authors in recent years have studied a number of distributed RC networks, in which energy is assumed to flow in one direction only[1-12]. The network structure consists of a thin film or layer of a dielectric material between a thin film of a resistive material and a substrate of an ideal conductor and a distributed transmission line is chosen as the model for its analysis. Network functions like the short circuit admittances of these 3-terminal networks have been derived for uniform lines and many special cases of nonuniform lines[1,2,3,9,10], thus providing us with a catalog of many networks with known behavior. Synthesis procedures to realize specified frequency characteristics by means of uniformly distributed RC networks have been recently advanced[11]. 'Structural topology' instead of component topology was considered to be the major approach to obtain different frequency response. Thus it has been said "---there is reason to believe that the generation of subnetworks may ultimately replace present synthesis procedures of generating network functions by connecting separate components".#

---

* This work was supported by NSF Grant No. GP-3017

W.W.Happ and G.C.Riddle "Limitations of Film Type Circuits Consisting of Resistive and Capacitive Layers" IRE International Convention Record,Part 6,1961. p.142.

A review of the transfer characteristics of uniform as well as nonuniform distributed RC networks under all terminal configurations and boundary conditions, that have been extensively discussed in the literature, shows the obvious result that such double-layer networks can give rise to either low-pass or high-pass frequency response only. It was Harris[†] who had originally suggested that a high-Q notch filter characteristic can be obtained by connecting a single resistance in series with a low-pass RC ladder network and Hager[6]constructed a distributed RC network instead of the lumped ladder network. Several authors who have studied special nonuniform distributed RC networks have added a discussion of the notch filter characteristics obtained when such networks are connected in series with a single lumped or diffused resistance. And a structure that has been commonly used for both theoretical analysis and practical fabrication by thin film and integrated circuit technology, to obtain a band-pass characteristic has embedded such a notch filter network in the feedback path of a high-gain amplifier.

PROPERTIES OF THE DOUBLE-KELVIN RC LINE

The purpose of this paper is to describe the properties of a four-layer distributed RC network which can be made to exhibit either low-pass, high-pass or band-pass frequency characteristics by imposing proper terminal conditions on its five terminals. The versatality of getting these three types of frequency response from the same basic network structure shown in Fig.1 and its suitability for practical fabrication by present-day technology are not so well known.
An incremental section of the transmission line that represents the distributed RC network is shown in Fig.2. The short circuit admittance matrix of the transmission line represented as a 5-terminal network element in Fig.3, is given by the following equation:

$$\begin{vmatrix} I_{1(0)} \\ I_{1(L)} \\ I_{2(0)} \\ I_{2(L)} \end{vmatrix} = \begin{vmatrix} y_{11} & y_{11'} & y_{12} & y_{12'} \\ y_{1'1} & y_{1'1'} & y_{1'2} & y_{1'2'} \\ y_{21} & y_{21'} & y_{22} & y_{22'} \\ y_{2'1} & y_{2'1'} & y_{2'2} & y_{2'2'} \end{vmatrix} \begin{vmatrix} V_{1(0)} \\ V_{1(L)} \\ V_{2(0)} \\ V_{2(L)} \end{vmatrix}$$

where

$$u = \sqrt{s r_1 c_1 L^2} \quad ; \quad \frac{r_2}{r_1} = n \quad ; \quad \frac{c_2}{c_1} = m$$

$$\tau = \frac{1}{2}\left(1 + \frac{r_2}{r_1} + \frac{r_2 c_2}{r_1 c_1}\right)$$

$$A = \sqrt{\tau + \sqrt{\tau^2 - nm}} \quad ; \quad \mathcal{K}_1 = \frac{1}{(1-B^2)}$$

$$B = \sqrt{\tau - \sqrt{\tau^2 - nm}} \quad ; \quad \mathcal{K}_2 = n\mathcal{K}_1$$

[†] G.R.Harris "Bridged Reactance-Resistance Networks", Proc.IRE, vol.37, August 1949.pp.882-887.

$$Y_{10} = \frac{Au}{r_1 L}\frac{(1-B^2)}{(A^2-B^2)} \quad ; \quad Y_{20} = \frac{Bu}{r_2 L}\frac{(1-B^2)}{(A^2-B^2)}$$

$$Y_{11} = Y_{1'1'} = Y_{10}\coth(Au) + \mathcal{X}_2^2 Y_{20}\coth(Bu)$$

$$Y_{22} = Y_{2'2'} = Y_{10}\mathcal{X}_1^2\coth(Au) + Y_{20}\coth(Bu)$$

$$Y_{1'1} = Y_{11'} = -Y_{10}\,\mathrm{cosech}(Au) - \mathcal{X}_2^2 Y_{20}\,\mathrm{cosech}(Bu)$$

$$Y_{2'2} = Y_{22'} = -Y_{10}\mathcal{X}_1^2\,\mathrm{cosech}(Au) - Y_{20}\,\mathrm{cosech}(Bu)$$

$$Y_{2'1'} = Y_{1'2'} = -Y_{10}\mathcal{X}_1\coth(Au) + \mathcal{X}_2 Y_{20}\coth(Bu) = Y_{21} = Y_{12}$$

$$Y_{2'1} = Y_{21'} = Y_{10}\mathcal{X}_1\,\mathrm{cosech}(Au) - \mathcal{X}_2 Y_{20}\,\mathrm{cosech}(Bu) = Y_{1'2} = Y_{12'}$$

r_1 and r_2 are resistances per unit length of the line

c_1 and c_2 are capacitances per unit length of the line

L = Length of the transmission line

Googe and Su[5]who called it a double-Kelvin transmission line studied only the low-pass frequency response of the voltage ratio transfer function $V_{1(L)}/V_{1(0)}$ under the boundary conditions $I_{1(L)} = I_{2(0)} = 0$; $I_{2(L)} = 0$ and its variation as the parameters $n = r_2/r_1$ and $m = c_2/c_1$ are varied. They also studied the interesting property that for certain values of n and m lying on a locus, the transfer function has a zero of transmission at a real (radian) frequency. However, Yanai et. al. had given in an earlier paper[12], not only the above set of boundary conditions but two more sets under which the basic network shown in Fig.1 exhibits a low-pass frequency response: (a) the voltage ratio $V_{1(L)}/V_{1(0)}$ when $V_{2(0)} = 0 = V_{2(L)}$; $I_{1(L)} = 0$ i.e.when terminals 2 and 2' are connected to ground and (b) the voltage ratio $V_{2(L)}/V_{1(0)}$ when $I_{2(0)} = I_{1(L)} = 0$; $I_{2(L)} = 0$ i.e.when terminals 2 and 1' are open have low-pass frequency response characteristic. Also in case (a), the voltage ratio transfer function is shown to possess transmission zeros at real frequencies when proper values of n and m are chosen. The authors state that the network has a steeper attenuation than a double layer RC line. They have presented other sets of boundary conditions under which the same network of Fig.1 exhibits high-pass, all-pass and also band-pass characteristics. But it has been found impossible to obtain an all-pass frequency response from the network under the conditions given by the authors.

BAND-PASS FREQUENCY RESPONSE

In the following section, our results of an extensive study of only the band-pass characteristics are summarized. Figure 4 shows the 2-port network when terminals 1' and 2 of the network are connected to ground terminal. The short circuit admittance matrix of the 2-port is obtained from the previous equation by deleting the second and third rows and also columnsof the matrix.

$$\begin{vmatrix} I_{1(0)} \\ I_{2(L)} \end{vmatrix} = \begin{vmatrix} Y_{10}\coth(Au) + Y_{20}\mathcal{X}_2^2\coth(Bu) & Y_{10}\mathcal{X}_1\,\mathrm{cosech}(Au) - \mathcal{X}_2 Y_{20}\,\mathrm{cosech}(Bu) \\ Y_{10}\mathcal{X}_1\,\mathrm{cosech}(Au) - \mathcal{X}_2 Y_{20}\,\mathrm{cosech}(Bu) & Y_{10}\mathcal{X}_1^2\coth(Au) + Y_{20}\coth(Bu) \end{vmatrix} \begin{vmatrix} V_{1(0)} \\ V_{2(L)} \end{vmatrix}$$

The magnitude and phase response of the open circuit voltage ratio

$$T_{21} = \frac{\mathcal{X}_2 Y_{20} \operatorname{cosech}(Bu) - Y_{10}\mathcal{X}_1 \operatorname{cosech}(Au)}{Y_{20} \coth(Bu) + \mathcal{X}_1^2 Y_{10} \coth(Au)}$$

has been calculated for several sets of the parameters n and m by using a digital computer. Figure 5 shows a few magnitude response curves for a fixed value of n = 1 and different values of m. All these curves coincide at low frequencies on an asymptotic slope of 10 db.per decade. As m decreases, both the center frequency ω_c and the gain at ω_c of the band-pass curves increases and the curves become less symmetric about their center frequency. However as m is decreased below some value depending on the value of n kept fixed, the increase in the center frequency becomes smaller and reaches a value which may be considered to be a lower limit for practical design. The phase response curves have the shape of a maximally flat low-pass characteristic and are nearly linear over the pass-band. When n is decreased keeping m fixed, the center frequency ω_c increases again; but the gain at ω_c remains constant, as seen from the curves in Fig.6. Changing the parameters r_1, r_2, c_1, c_2 and L of an integrated or a thin film circuit after their fabrication is difficult but the ratio of the resistances and capacitances per unit length can be controlled within a small percentage, during the diffusion process. It is however possible to change the capacitances c_1 and c_2 individually, by adjusting the reverse bias voltage applied to the two p-n junctions of an integrated circuit, after it is fabricated. From these considerations, a more accurate control of the resistances r_1 and r_2 than that of the capacitances seems to be very desirable.

The band-pass network, no doubt, has a low value of Q but this can be improved by cascading a number of such networks. In order to find the loading effects, the o.c. impedances and the s.c. admittances of the network have been calculated as a function of frequency. It is found that the immittance levels at the two ports are not the same. The short circuit admittances have nearly a constant value over the pass-band but the open circuit impedances have a slightly drooping characteristics beyond the center frequency.

From the Nyquist plot of the transfer function, it is found that the band-pass network can be used to build notch filters and oscillators by connecting a resistance in series with this 2-port. Only the notch filter characteristics are shown in Fig.7 for different values of the series resistance R_s connected to a band-pass network having n=1, m=1 and r_1L =500. The maximum null appears about one decade beyond the center frequency of the band-pass network. A second null exists at a still higher value of frequency and for a lower value of R_s. As the value of R_s is increased, the depth of the null as well as the null frequency decreases and the slope of the low frequency asymptote also decreases. It has also been found that as m and/or n is decreased, the null frequency increases; thus tuning the capacitance ratio m by adjusting the reverse junction bias, provides us with some control on the null frequency also.

CONCLUSION

A more detailed investigation of the low-pass,high-pass and band-pass characteristics of the network than described here is being carried out by this author. The tranfer function obtained under the conditions claimed by Yanai et al to give an all-pass frequency response and those that would be realized under other sets of boundary conditions would also be studied from a theoretical interest. Even from a practical point of interest, the network shown in Fig.1 deserves more detailed investigation. Its structure is a better approximation to the component network elements that are built in monolithic integrated circuits. At present these elements -like the resistors and diodes- are fabricated by the same process that is used to build transistors,as seen from the cross section of a typical monolithic integrated circuit shown in Fig. 8. This process involves diffusion of two layers by impurities, giving rise to a structure having distributed capacitances between two p-n junctions that are separated by two resistive layers. In the resistors,both junctions are highly reverse biassed and the capacitances are considered as parasitic elements. But a different attitude that can be taken is to study how these capacitances influence the frequency characteristics and how they can be controlled during fabrication to attain such values that the frequency response of the resulting morphology is of desirable nature. A large number of such network elements that approximate the structure shown in Fig.1 can be grown in a single chip and by properly interconnecting the terminals and connecting the resulting two-port networks,it is possible to realize low-pass,high-pass and band-pass frequency response characteristics from one common type of distributed RC network. Of course, by means of thin film technology,the network can be more easily fabricated and its parameters more accurately controlled.

REFERENCES

1. Dutta Roy,S.C."Some Exactly Solvable Nonuniform RC Lines". IEEE Trans.on Circuit Theory,CT-12,March 1965.pp.141-142.

2. -------------"Matrix Parameters of Nonuniform Transmission Lines" IEEE Trans.on Circuit Theory, CT-12, March 1965.pp.142-143.

3. -------------"Theory of Exponentially Tapered RC Transmission Lines for Phase Shift Oscillators". Proc.IEE,vol. 110, October 1963. pp.1764-1770.

4. Edson, W.A."Tapered Distributed RC Lines for Phase Shift Oscillators Proc.IRE,vol.49, June 1961.pp.1021-1024.

5. Googe, J.M. and Su, K.L." The Double-Kelvin Transmission Line and Some Applications" IEEE Trans.on Circuit Theory, CT-11,September 1964. pp372-377.

6. Hager,C.K."Distributed Parameter Networks for Circuit Miniaturization" Proc.Elec.Compts.Conf.Philadelphia.Pa.May 1959. pp.195-203.

7. Happ,W.W."Distributed Parameter Circuits and Microsystems Electronics" Proc.NEC. vol.16,1960.pp.448-460.

8. Kaufman,W.M." Theory of a Monolithic Null Device and Some Novel Circuits" Proc.IRE,vol.48, September 1960, pp.1540-1545.

9. Kaufman,W.M. and Garrett,S.J."Tapered Distributed Filters" IEEE Trans.on Circuit Theory,CT-9,December 1962. pp.329-336.

10. Su,K.L."The Trigonometric RC Transmission Line" IEEE International Convention Record,Part II, 1963. pp.43-55.

11. Wyndrum,R.W."The Exact Synthesis of Distributed RC Networks". N.Y.U. Tech.Report No.400-76. May 1963.

12. Yanai,H.,Sugano,T.,Sasaki,H.,and Yoshida,M."Theory and Characteristics of Multilayer RC Distributed Constant Circuits" J.I.E.C.E.(Japan) vol.46, March 1963. pp.323-329.

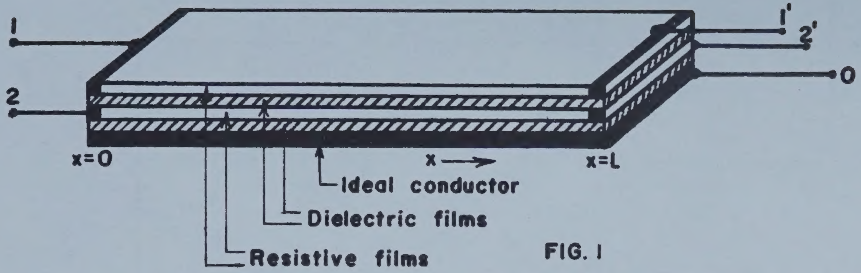

1
2
1'
2'
0
x=0
x=L
x ⟶
Ideal conductor
Dielectric films
Resistive films

FIG. 1

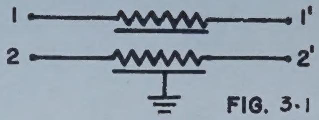


FIG. 3-1

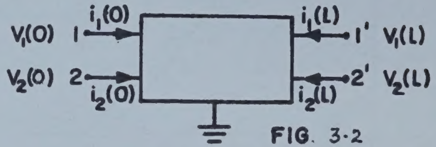

V₁(0) 1
V₂(0) 2
i₁(0)
i₂(0)
i₁(L)
i₂(L)
1' V₁(L)
2' V₂(L)

FIG. 3-2

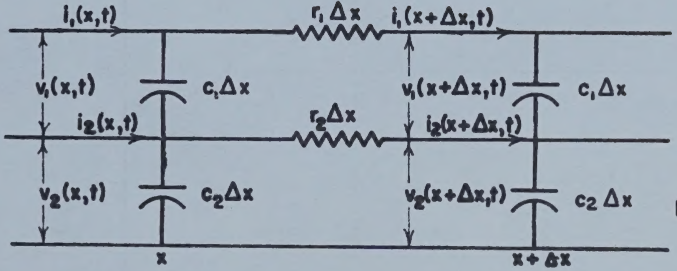

i₁(x,t)
r₁Δx
i₁(x+Δx,t)
v₁(x,t)
c₁Δx
v₁(x+Δx,t)
c₁Δx
i₂(x,t)
r₂Δx
i₂(x+Δx,t)
v₂(x,t)
c₂Δx
v₂(x+Δx,t)
c₂Δx
x
x+Δx

FIG. 2

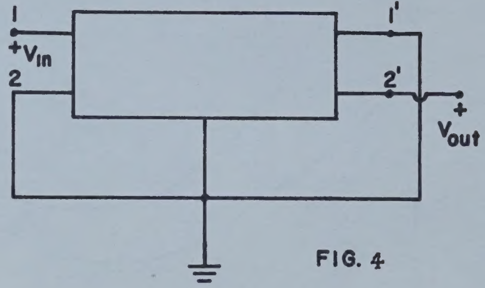

1
+V_in
2
1'
2'
+
V_out

FIG. 4

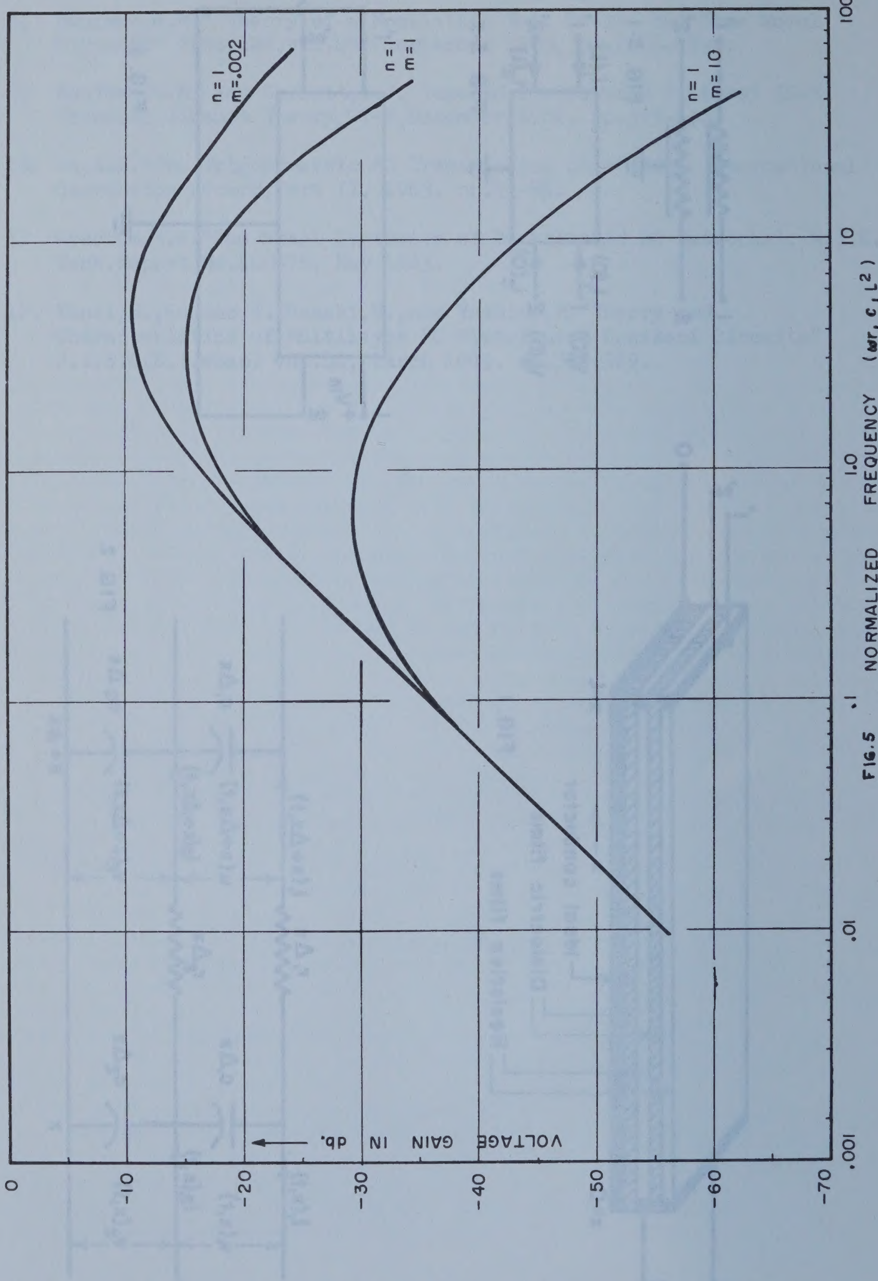

VOLTAGE GAIN IN db.
0
-10
-20
-30
-40
-50
-60
-70
.001
.01
.1
1.0
10
100
NORMALIZED FREQUENCY (ωr, c, L²)
n = 1 m = .002
n = 1 m = 1
n = 1 m = 10

FIG. 5

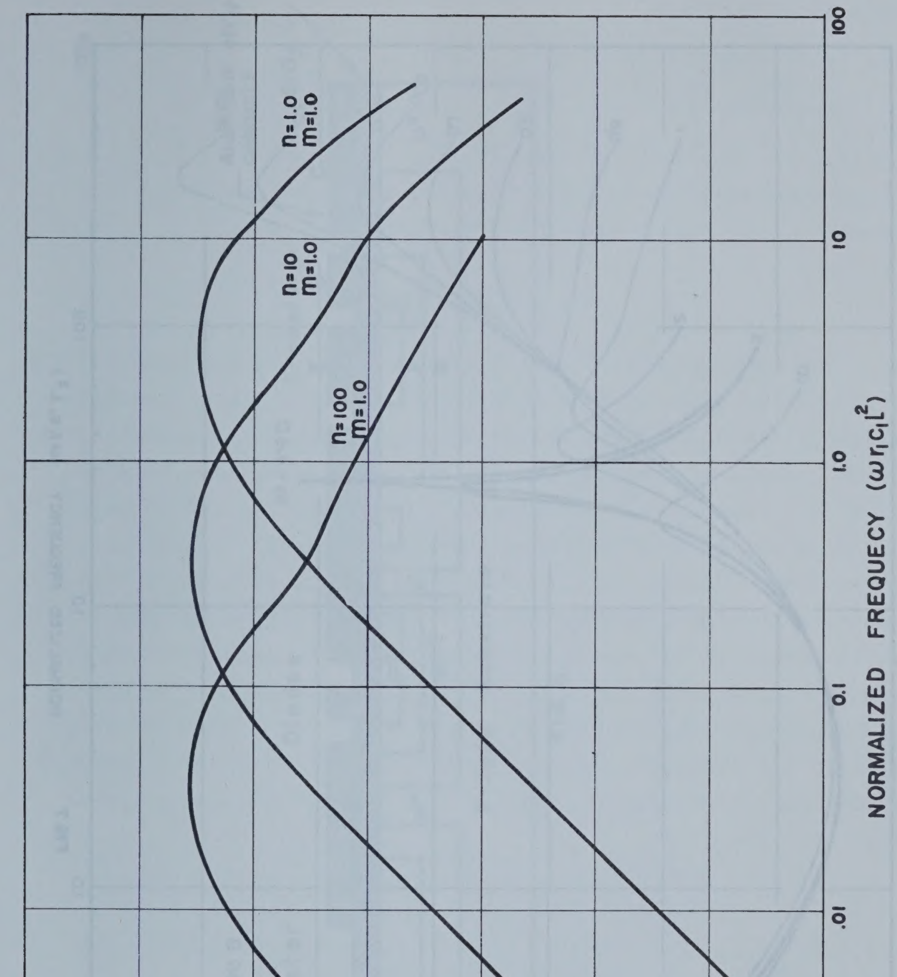

VOLTAGE GAIN IN db.
0
-10
-20
-30
-40
-50
-60
-70
.001
.01
0.1
1.0
10
100
NORMALIZED FREQUECY $(\omega r_1 c_1 l^2)$
n=1.0
m=1.0
n=10
m=1.0
n=100
m=1.0

FIG. 6

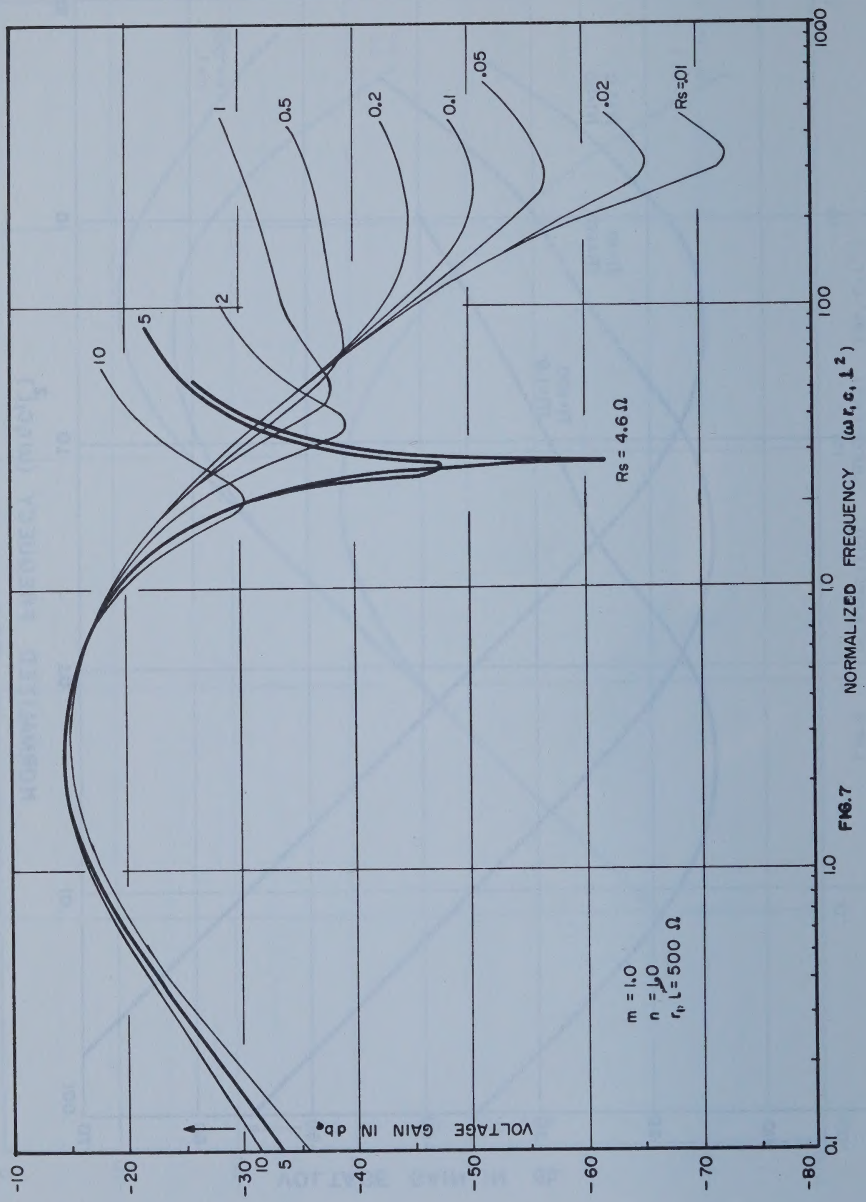

-10
-20
-30
-40
-50
-60
-70
-80
10
5
VOLTAGE GAIN IN db
0.1
1.0
1.0
100
1000
NORMALIZED FREQUENCY (ω r, c, l²)
10
5
2
1
0.5
0.2
0.1
.05
.02
Rs=.01
Rs = 4.6 Ω
m = 1.0
n = 1.0
r_p, l = 500 Ω

FIG.7

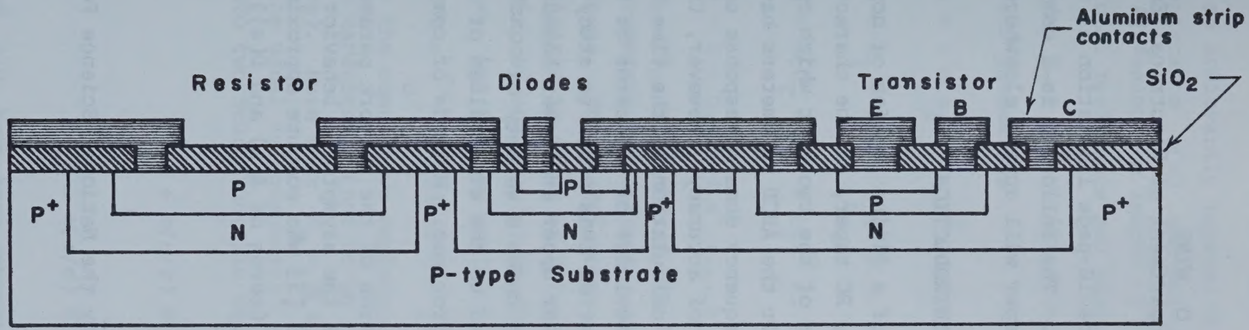

Resistor
Diodes
Transistor
Aluminum strip contacts
SiO2
E
B
C
P
P
P
P+
P+
P+
P+
N
N
N
P-type Substrate

FIG. 8.

DISTRIBUTION OF ZEROS OF ABCD PARAMETERS OF ARBITRARY RC TRANSMISSION LINES*

E.N. PROTONOTARIOS and O. WING
Columbia University, Department of Electrical Engineering
New York, N.Y.

(Note: The paper exceeds the 10-page limitation as stipulated in the Instructions to authors. The following is a comprehensive summary. The full-length paper will appear elsewhere.)

1. INTRODUCTION

The terminal behavior of a finite section of non-uniform RC transmission line, or simply RC taper, can be characterized in terms of the ABCD parameters of the two-port which represents the taper. Explicit formulas for the ABCD parameters have been derived [1] from which the frequency domain response can be calculated to any desired degree of accuracy. However, these formulas are not useful in the calculation of the time domain response, which requires the knowledge of the zeros of the ABCD parameters. It is the object of this paper to study the properties of these zeros, and their upper and lower bounds, for classes of RC tapers defined under a variety of constrains on the taper. The zeros are related to the eigenvalues of four properly formulated Sturm-Liouville problems. Methods of computation of the zeros are given.

Previous work on the zeros of the network parameters is that of Hellstrom [2] who studies the asymptotic behavior of the zeros and that of Kelly and Ghausi [3] who compute approximately the poles of transfer functions (zeros of A(s) and D(s)) of an exponentially tapered RC line.

---

* This work was supported by the National Science Foundation under NSF Grant G-1505.

2. DEFINITIONS

Consider an arbitrarily tapered RC-transmission line with per-unit length resistance $r(x)$ and per-unit length capacitance $c(x)$, $x \in [0, \ell]$, ℓ being the length of the line. Denote by R_T and C_T the total resistance and capacitance of the line, i.e. $R_T = \int_0^{\ell} r(x)dx$, $C_T = \int_0^{\ell} c(x)dx$. Consider the following two forms of the ABCD parameters

$$A(s) = 1 = a_1 s + a_2 s^2 + \ldots = \prod_{n=1}^{\infty} (1 + s/\alpha_n) \tag{1}$$

$$B(s) = R_T + b_1 s + b_2 s^2 + \ldots = \prod_{n=1}^{\infty} (1 + s/\beta_n) \tag{2}$$

$$C(s) = s(C_T + c_1 s + c_2 s^2 + \ldots) = \prod_{n=1}^{\infty} (1 + s/\gamma_n) \tag{3}$$

$$D(s) = 1 + d_1 s + d_2 s^2 + \ldots = \prod_{n=1}^{\infty} (1 + s/\delta_n) \tag{4}$$

Where $-\alpha_n$, $-\beta_n$, $-\gamma_n$ and $-\delta_n$ $(n = 1, 2, \ldots)$ are the zeros of the ABCD parameters respectively, arranged in order of increasing magnitude. The coefficients a_n, b_n, c_n, d_n $(n = 1, 2, \ldots)$ are given explicitly in terms of integrals of $r(x)$ and $c(x)$ in Ref 1. Let $V(x, s)$ and $I(x, s)$ be the voltage and current respectively at x. Let two new variables be introduced

$$R = \int_0^x r(x')dx', \quad C = \int_0^x c(x')dx'$$

so that R is the cumulative resistance along the line and C is the cumulative capacitance (not to be confused with $C(s)$ of the ABCD parameters). Note that $R_T = R(\ell)$, $C_T = C(\ell)$. It is easy to show that the usual telegraphists' equations on V and I are transformed into two Sturm-Louville equations given below

$$\frac{d^2V}{dR^2} + s(c/r)\, V = 0 \tag{5}$$

$$\frac{d^2I}{dC^2} + s(r/c)\, I = 0 \tag{6}$$

where c/r and r/c have been expressed as functions of R and C respectively. Either may have discontinuities or contain impulses (corresponding to lumped elements). It is convenient to define a function $\sigma(R)$ such that

$$C = \sigma(R) = \sigma(R\text{-}0) \quad \text{for} \quad 0 \leq R < R_T$$

$\sigma(R)$ implicitly specifies the shape of the taper. RC-transmission lines with the same function $C = \sigma(R)$, $R \in (0, R_T)$ have identical electrical characteristics at their terminals i.e. they are equivalent. Note that $\sigma'(R) = d\sigma/dR = c/r$. $\sigma(R)$ is a non-decreasing function and therefore it has always a unique inverse i.e. $R = \tau(C)$, $C \in (0, C_T)$ with $\tau'(c) = d\tau/dC = r/c$. Note that lumped shunt capacitances correspond to jumps of $\sigma(R)$. (Fig. 1)

3. THE FOUR STURM-LIOUVILLE PROBLEMS

It is easy to see that the zeros of $B(s)$ and $C(s)$ are the eigenvalues of the Sturm-Liouville equations (5) and (6) respectively with zero boundary conditions at both ends i.e. $V(0) = V(R_T) = 0$ for (5) and $I(0) = I(C_T) = 0$ for (6). The zeros of $A(s)$ and $D(s)$ are easily identifiable with the odd order eigenvalues of the Sturm-Liouville equations (5) and (6) respectively but with doubled intervals of definition, i.e. $(0, 2R_T)$ for (5) and $(0, 2C_T)$ for (6). The functions c/r and r/c, in the equations are symmetrically extended in the intervals $(R_T, 2R_T)$ and $(C_T, 2C_T)$ respectively i.e.

$$c/r = \sigma'(R) = \sigma'(2R_T\text{-}R) \quad \text{for} \quad R \in (R_T, 2R_T) \quad \text{and}$$

$$r/c = \tau'(C) = \tau'(C_T\text{-}C) \quad \text{for} \quad C \in (C_T, 2C_T).$$

A variational formula may now be given for β_1, as follows

$$\beta_1 = \min_{v(R) \in \Gamma^1} \left(\int_0^{R_T} v'^2 dR \Big/ \int_0^{R_T} v^2 \, \sigma'(R) \, dR \right) \qquad (7)$$

where Γ^1 is the class of continuously differentiable functions in the interval (o, R_T) with $v(0) = v(R_T) = 0$. Using this formula we can obtain a numerical approximation to β_1 by substituting for $v(R)$ in (7) functions $\in \Gamma^1$ which resembles the eigenfunction. For the higher order zeros of $B(s)$ we refer to the developments in Ch. VI of Courant and Hilbert Ref. 4 which examines in detail the variational formulas for the eigenvalues.

4. BOUNDS ON THE ZEROS OF THE ABCD PARAMETERS

4.1 If for some arbitrarily tapered RC-transmission line nothing is specified, apart from its total resistance R_T and its total capacitance C_T, then for the first zero $-\beta_1$ of B(s) it can be asserted that:

$$\beta_1 \geq \frac{4}{R_T C_T} \tag{8}$$

This inequality is the best possible, i.e. there exist tapers for which the value of β_1 is arbitrarily close to $4/R_T C_T$. In fact the equality is attained for the T-network in Fig. 2. For the higher order zeros of B(s) and for the zeros of A, C and D parameters we have the following "best possible" inequalities

$$\alpha_n,\ \delta_n \geq \frac{(2n-1)^2}{R_T C_T} \ ; \quad \beta_n,\ \gamma_n \geq \frac{4n^2}{R_T C_T},\ n = 1, 2...$$

4.2 As a second case suppose that the only thing specified is $\Delta = \int_0^{\ell} \sqrt{rc}\, dx \leq R_T C_T$. Without any additional information about the taper it is not difficult to construct examples showing that the product $\beta_1 \Delta^2$ can get arbitrarily small. However, with the additional requirement that c/r is a non-negative monotonic function of R, $\in L^1\, [0, R_T]$ we can assert that :

$$\beta_1\ \Delta^2 \geq \Pi^2/4 \quad \text{and} \quad \gamma_1\ \Delta^2 \geq \Pi^2/4$$

Note that those inequalitites are the best possible. [7] In general for $\sigma'(R) \in L^1\, [0, R_T]$ and if the interval $[0, R_T]$ can be decomposed into k + 1 intervals in which $\sigma'(R) = c/r$ is monotonic then the zeros obey the following inequalities

$$\alpha_n,\ \delta_n > (n-k-1)^2\ \Pi^2/4\Delta^2; \quad \beta_n,\ \gamma_n > (n-k)^2\ \Pi^2/4\Delta^2 \text{ with } n > k$$

4.3a Let $R_T > 0$, $C_T > 0$, $K_M > 0$, $K_m > 0$ be numbers subject to the condition

$$K_m^{\ 2} \leq R_T/\ C_T \leq K_M^{\ 2}$$

Consider the class of RC-transmission lines with total series resistance R_T, total shunt capacitance C_T and

$K_m \leq \sqrt{r(x)/c(x)} \leq K_M$ for $0 \leq x \leq \ell$. Denote this class of tapers by: $T(R_T, C_T, K_M, K_m)$. Since $\sigma'(R) = c/r$ we have:

$1/K_M^2 \leq \sigma'(R) \leq 1/K_m^2$ i.e. all the points of the plot of the function $C = \sigma(R)$ must lie inside the parallelogram whose sides have slopes $1/K_m^2$ and $1/K_M^2$ and two of its vertices coincide with the origin and the point (R_T, C_T) respectively. (Fig. 3). The following theorem is deducible from Krein's work [6].

Theorem For an RC-transmission line $\in T(R_T, C_T, K_M, K_m)$ the maximum and minimum values of the ABCD parameters are given by the following relations:

$$\text{(i)} \quad \max \alpha_n(\sigma) = \max \delta_n(\sigma) = \left[\frac{(2n-1)K_M}{Q} \varphi\left(\frac{K_M}{K_m}, \frac{q}{Q} \right)\right]^2$$

$$\min \alpha_n(\sigma) = \min \delta_n(\sigma) = \left[\frac{(2n-1)K_m}{q} \varphi\left(\frac{K_m}{K_M}, \frac{Q}{q} \right)\right]^2$$

$$\text{(ii)} \quad \max \beta_n(\sigma) = \max \gamma_n(\sigma) = \left[\frac{2nK_M}{Q} \varphi\left(\frac{K_m}{K_M}, \frac{q}{Q} \right)\right]^2$$

$$\min \beta_n(\sigma) = \min \gamma_n(\sigma) = \left[\frac{2nK_m}{q} \varphi\left(\frac{K_m}{K_M}, \frac{Q}{q} \right)\right]^2$$

where, with reference to Fig. 3:

$$q = K_m^2 \frac{K_M^2 C_T - R_T}{K_M^2 - K_m^2}, \quad Q = K_M^2 \frac{R_T - K_m^2 C_T}{K_M^2 - K_m^2}$$

and $\varphi(p, \theta)$ $(0 \leq p, \theta < \infty)$ is the least positive root of the equation:

$$\tan \varphi \ \tan (p \ \theta \ \varphi) = p$$

(thus $0 < \varphi(p, \theta) < \frac{\Pi}{2}$).

The lower bound of $\alpha_1(\sigma)$ and the upper bound $\beta_1(\sigma)$ are attained "uniquely" for $\sqrt{r/c}$ with a bang-bang x variation i.e for

$$\sqrt{\frac{r(R)}{c(R)}} = \begin{cases} K_M & R \in (0, Q) \\ K_m & R \in (Q, R_T) \end{cases}$$

note that, for this case, the plot of the function $C = \sigma(R)$ is the broken line 0 Π P of Fig. 3.

Similar results hold for α_n, δ_n, and γ_n. If a_1 is known, using the obtained bound and the relation

$$\sum_{i=1}^{\infty} 1/\alpha_i = a_1$$

we can find a sharper result for the interval where α_1 must lie, i.e.

$$\frac{1}{a_1 - \dfrac{0.2337\,Q^2}{[\,K_M\,\varphi(\frac{K_M}{K_m},\frac{q}{Q})\,]^2}} \le \alpha_1 \le \frac{1}{a_1 - \dfrac{0.2338\,q^2}{[\,K_m\,\varphi(\frac{K_m}{K_M},\frac{Q}{q})\,]^2}}$$

where $0.2337 \cong \frac{\Pi^2}{8} - 1 = \sum_{n=2}^{\infty} \frac{1}{(2n-1)^2}$

4.3b Two special cases are studied in the paper. In case 1, $K_M = \infty$, so that lumped series resistances are allowed; this class of tapers is denoted by $T(R_T, C_T, K_m)$. In case 2, $K_m = 0$, so that lumped shunt capacitances are allowed; we denote this class by $T(R_T, C_T, K_M)$. We will include here the upper and lower bounds only.

Theorem 1. For $\sigma(R) \in T(R_T, C_T, K_m)$ we have:

(i) $\max \alpha_n = [\,\frac{(2n-1)\,\Pi}{2C_TK_m}\,]^2$

$$\min \alpha_n = [\,\frac{(2n-1)}{C_TK_m}\,\psi(\frac{C_TK_m^2}{R_T})\,]^2 > \frac{(2n-1)^2}{R_TC_T} \cdot \frac{1}{1-(1-\frac{4}{\Pi^2})\frac{C_TK_m^2}{R_T}}$$

(ii) $\max \beta_n = [\,\frac{n\Pi}{C_TK_m}\,]^2$

$$\min \beta_n = [\,\frac{2n}{C_TK_m}\,\varphi(\frac{C_TK_m^2}{R_T})\,]^2 > \frac{4n^2}{R_TC_T} \cdot \frac{1}{1-(1-\frac{4}{\Pi^2})\,\frac{C_TK_m^2}{R_T}}$$

where $\psi(\theta)$ is the least positive root of the equation

$$\psi \tan \psi = \frac{\theta}{1-\theta}\,.$$

Theorem 2 For $\sigma(R) \in T(R_T, C_T, K_M)$ we have:

(i) $\max \alpha_n = [\,\frac{(2n-1)\,\Pi\,K_M}{2R_T}\,]^2$

$$\min \alpha_n = \left[\frac{(2n-1)K_M}{R_T} \psi \left(\frac{R_T}{C_T K_M^2} \right) \right]^2 > \frac{(2n-1)^2}{R_T C_T} \cdot \frac{1}{1-(1-\frac{4}{\Pi^2}) \frac{R_T}{C_T K_M^2}}$$

(ii) $$\max \beta_n = \left[\frac{n \Pi K_M}{R_T} \right]^2$$

$$\min \beta_n = \left[\frac{2n K_M}{R_T} \psi \left(\frac{R_T}{C_T K_M^2} \right) \right]^2 > \frac{4n^2}{R_T C_T} \cdot \frac{1}{1-(1-\frac{4}{\Pi^2}) \frac{R_T}{C_T K_M^2}}$$

Where $\psi(\theta)$ is defined as in Theorem 1. Note that the upper and lower bounds of γ_n and δ_n are equal to those of β_n and α_n respectively in both theorems above.

5. ASYMPTOTIC BEHAVIOR OF THE ZEROS

Using a theorem by Titchmarsh [8] for the distribution of zeros of a special form of entire functions we can prove that the poles $-\sigma_n$ of any transfer function (therefore for the zeros of ABCD in particular) of an RC-transmission line terminated in a lumped RC-impedance satisfy the following asymptotic relation:

$$\sigma_n \sim (n\Pi/\Delta)^2 \quad \text{for} \quad n \to \infty$$

where $$\Delta = \int_0^{\ell} \sqrt{rc}\, dx \, .$$

For the last statement the only assumption for $r(x)$, $c(x)$ is that they are non-negative integrable, in the distribution sense, functions of x, i.e. they may contain discontinuities and impulses.

6. COMPUTATION OF THE ZEROS

The sum of negative powers of the zeros of the ABCD parameters can be expressed in terms of the coefficients of the series (1), (2), (3) and (4), page 2.

Specifically we have:

$$\sum_{n=1}^{\infty} \frac{1}{\alpha_n} = a_1 \, , \quad \sum_{n=1}^{\infty} \frac{1}{\alpha_n^2} = a_1^2 - 2a_2 , \quad \sum_{n=1}^{\infty} \frac{1}{\alpha_n^3} = a_1^3 - 3a_1 a_2 + 3a_3 \quad (9)$$

Similar relations hold for the zeros of $B(s)$, $C(s)$ and $D(s)$.

An approximate method for the computation of the zeros of the ABCD parameters can be based on the equation (9). Taking under consideration that, begining with a sufficiently large subscript all zeros may be regarded as known, i.e. equal to their asymptotic values then an algebraic system of equations is obtained from (9), for the first few zeros.

To be specific suppose that we want to compute the first few zeros of A(s). The easiest asymptotic estimate of α_n, for large n, is $\alpha_n \approx (2n-1)^2 \Pi^2/\Delta^2$. Suppose that this asympotic formula gives a "good" approximate value of α_n beginning with n=4. Then we will have for the three first zeros of A(s) the following algebraic system:

$$\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \frac{1}{\alpha_3} = a_1 - \left(\frac{2\Delta}{\Pi}\right)^2 \zeta^* (2,4)$$

$$\frac{1}{\alpha_1^2} + \frac{1}{\alpha_2^2} + \frac{1}{\alpha_3^2} = a_1^2 - 2a_2 - \left(\frac{2\Delta}{\Pi}\right)^4 \zeta^*(4, 4)$$

$$\frac{1}{\alpha_1^3} + \frac{1}{\alpha_2^3} + \frac{1}{\alpha_3^3} = a_1^3 - 3a_1a_2 + 3a_3 - \left(\frac{2\Delta}{\Pi}\right)^6 \zeta^* (6, 4)$$

where $\zeta^* (s, \nu)$ is a modified Riemann's Zeta function defined as follows:

$$\zeta^* (s, \nu) = \sum_{n=\nu}^{\infty} \frac{1}{(2n-1)^s} \qquad (\nu \text{ positive integer}).$$

Note that $\zeta^* (2, 4) \cong 0.08589$, $\zeta^* (4, 4) \cong 0.0007323$ and $\zeta^* (6, 4) = 0.00001134$.

REFERENCES

1. E.N. Protonotarios and O. Wing, Delay and Rise Time of Arbitrarily Tapered RC-transmission Lines, 1965 IEEE International Convention Record, Part 7, pp. 1-6.
2. M.J. Hellstron, The Distributed rc Network, Ph.D. Dissertation, Polytechnic Institute of Brooklyn, June 1963.
3. J. Kelly and M. Ghausi, On the Dominant Poles of the Distributed RC Networks, Technical Report 400-92, New York University, March 1964.

4. Courant and Hilbert, Methods of Mathematical Physics, Vol. I, Interscience Publishers, Inc., New York, 1953.
5. P. Hartman and A. Winger, On an Oscillation Criterion of Liapounoff, Amer. J. Math., vol. 73 (1951), pp. 885-890.
6. M. G. Krein, On Certain Problems on the Maximum and Minimum of Characteristic Values and on Lyapnov Zones of Stability, Prikel. Mat. Meh. 15, 323-348 (1951) and in translation: AMS Translations S.2, Vol. 1, Providence, R.I., 1955, pp. 112-163.
7. Z. Nehari, Some Eigenvalue Estimates, Journal d' Analyse Mathematique, Vol. 7, Jerusalem 1959-60, pp. 79-88.
8. E.C. Titchmarsh, On Integral Functions with Yeal Negative Zeros, Proc. London Math. Soc. (2), 26(1927), pp. 185-200.

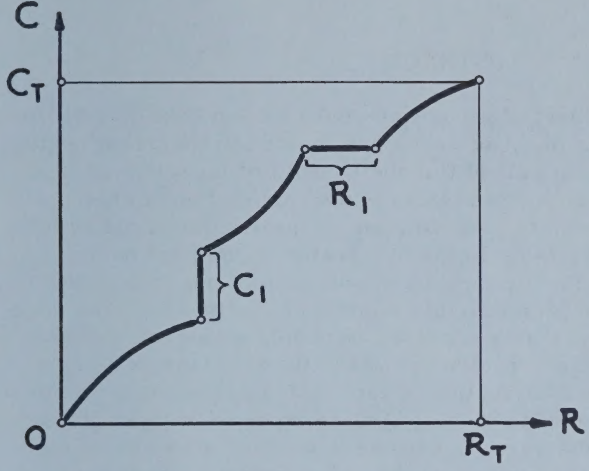


FIG. 1

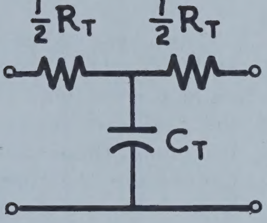

$\frac{1}{2}R_T$
$\frac{1}{2}R_T$
C_T

FIG. 2

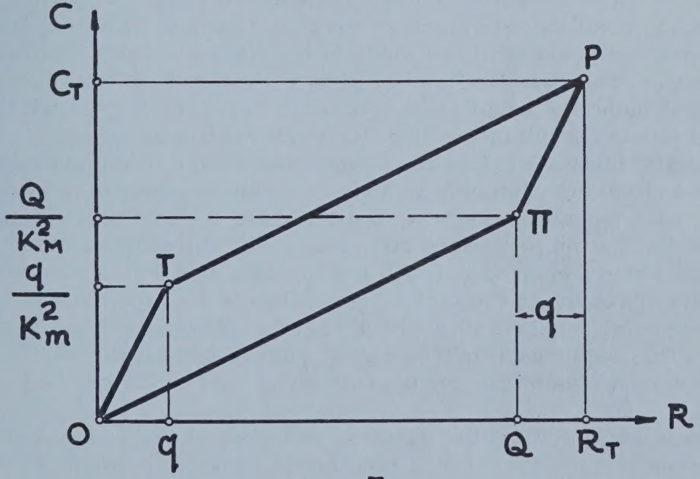


FIG. 3

THE GENERALIZATION OF THE MIYATA REALIZATION PROCEDURE

HYONG KAP KIM
The University of Manitoba, Winnipeg, Canada

ABSTRACT

The zeros of the even part of a positive real function exhibit quadrantal symmetry. If some of these zeros are located in a certain region of the complex plane, not all of the coefficients of the numerator of the real part are positive. The present work applies the extension of Richards' theorem in order to remove a set of quadrantal zeros which causes the negative coefficients in the numerator of the real part. Thus it forces the given function toward a subclass of positive real functions which is amenable to the Miyata method of synthesis. The necessary manipulation to attain this subclass is purely algebraic and simple. The number of cycles required to reach the subclass is determined by the density of the distribution of the quadrantal zeros in the prescribed region of the complex plane. Unlike other techniques the closeness of the zeros of the even part to the imaginary axis has no adverse effect. Proof of the method as well as illustrative examples are provided.

I. INTRODUCTION

Application of Richards' theorem(1) enabled Bott and Duffin (2) to prove that any positive real function could be realized as an impedance of a network composed of three kinds of ordinary elements, namely, R, L, and C. The Bott-Duffin method is very significant in that the use of mutual inductive coupling is avoided. However it generally requires a large number of elements than Brune(3) realization.

Miyata(4) method applies the proper separation of the even part of the given function and synthesis is done in terms of properties of the even part of each separation. Where applicable, it results in a relatively simple realization procedure and, except for the simplest functions, is inherently more economical of elements than the Bott-Duffin method. The only drawback is the lack of generality in its application, i. e. only a very special subclass of positive real functions is subject to the method. This subclass requires rather severe constraints on the location in the s-plane of the zeros of the even part of the positive real functions.

Balabanian(5) defined the subclass in terms of the location of its even part zeros in s-plane. For a function of rank 2n to belong to the subclass, the even part zeros of quadrantal symmetry are not to be located closer to the imaginary axis than an angle of $\pi/2n$ radians. If the even part zeros are not located closer to the imaginary axis than an angle of $\pi/4$ radians, the function is necessarily of the subclass. In order to force the given function to belong to the subclass one may apply the "Method of Order Elevation"(4). It multiplies the numerator and the denominator by proper surplus factors so as to eventually yield the subclass. This process, however, gives rise to the elevation of the rank by two, thus resulting in an increase of complexity. Furthermore, the number of surplus factors grows without limit as the even part zeros move colser toward the imaginary axis.

Present paper develops a new approach which systematically reduces the complexity and, at the same time, forces the given function directly toward the subclass. The basic operation involved is the removal of an unfavorable set of the quadrantal zeros. Unlike the other techniques, the closeness of zeros to the imaginary axis has no adverse effects.

II. REMOVAL OF AN UNFAVORABLE SET OF THE EVEN PART ZEROS

Let Z(s) be a positive real function of minimum reactance and minimum susceptance. Applying the Richards' theorem twice successively, one obtains

$$P(s) = \frac{aZ(s) - sZ(a)}{aZ(a) - sZ(s)} \tag{1}$$

$$Q(s) = \frac{bP(s) - sP(b)}{bP(b) - sP(s)} \tag{2}$$

Substitution of (1) into (2) gives

$$F(s) = Z(a)P(b)Q(s)$$

$$= \frac{\left[1+\frac{aZ(b)-bZ(a)}{ab(aZ(a)-bZ(b)}s^2\right]Z(s)-\left[\frac{(a^2-b^2)Z(a)Z(b)}{ab(aZ(a)-bZ(b))}s\right]}{\left[1+\frac{aZ(a)-bZ(b)}{ab(aZ(b)-bZ(a)}s^2\right]-\left[\frac{a^2-b^2}{ab(aZ(b)-bZ(a))}s\right]Z(s)} \tag{3}$$

It is obvious that the new function F(s) is a positive real function if both a and b are positive and real. If we let a and b represent a pair of complex conjugate zeros s_1 and $\bar{s}_1$ of Fig. 1, the following equations are obtained

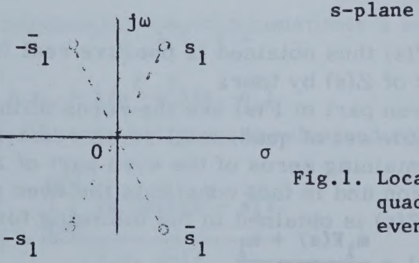


Fig.1. Location of a set of quadrantal zeros of the even part of Z(s)

$$
\begin{aligned}
a &= s_1 = \sigma_1 + j\omega_1 \\
b &= \bar{s}_1 = \sigma_1 - j\omega_1
\end{aligned} \tag{4}
$$

and

$$
\begin{aligned}
Z(a) &= U_1 + jV_1 \\
Z(b) &= U_1 - jV_1
\end{aligned} \tag{5}
$$

Substitution of eqs. (4) and (5) into (3) gives

$$F(s) = \frac{m_2 Z(s) - n_1}{m_1 - n_2 Z(s)} \tag{6}$$

where

$$
\begin{aligned}
m_1 &= As^2 + 1 \\
m_2 &= Cs^2 + 1 \\
n_1 &= Bs \\
n_2 &= Ds
\end{aligned} \tag{7}
$$

and

$$
\begin{aligned}
A &= \frac{\omega_1 U_1 + \sigma_1 V_1}{(\sigma_1^2 + \omega_1^2)(\omega_1 U_1 - \sigma_1 V_1)} \\
B &= \frac{2\sigma_1 \omega_1 (U_1^2 + V_1^2)}{(\sigma_1^2 + \omega_1^2)(\omega_1 U_1 + \sigma_1 V_1)} \\
C &= \frac{\omega_1 U_1 - \sigma_1 V_1}{(\sigma_1^2 + \omega_1^2)(\omega_1 U_1 + \sigma_1 V_1)} \\
D &= \frac{2\sigma_1 \omega_1}{(\sigma_1^2 + \omega_1^2)(\omega_1 U_1 - \sigma_1 V_1)}
\end{aligned} \tag{8}
$$

Hazony proved that F(s) thus obtained is positive real function whose rank is less than that of Z(s) by four.

The zeros of the even part of F(s) are the zeros of the even part of Z(s) with an unfavorable set of quadrantal zeros removed. Now it will be proved that the remaining zeros of the even part of Z(s) are undisturbed by this operation and in fact constitute the even part zeros of F(s). From eq. (6) Z(s) is obtained in the following form.

$$Z(s) = \frac{m_1 F(s) + n_1}{m_2 + n_2 F(s)} \tag{9}$$

Separating the even term and odd term of the numerator and denominator respectively, one obtains

$$Z(s) = \frac{\left[m_1 EvF(s)\right] + \left[n_1 + m_1 OdF(s)\right]}{\left[m_2 + n_2 OdF(s)\right] + \left[n_2 EvF(s)\right]} = \frac{m_1' + n_1'}{m_2' + n_2'}$$

$$EvZ(s) = \frac{m_1' m_2' - n_1' n_2'}{\left|m_2' + n_2'\right|^2} = \frac{m_1 m_2 - n_1 n_2}{\left|m_2' + n_2'\right|^2} EvF(s) \tag{10}$$

The numerator of eq. (10) is

$$\left|m_2' + n_2'\right|^2 = m_2^2 + 2m_2 n_2 OdF(s) + n_2^2 \left[OdF(s)\right]^2 - n_2^2 \left[EvF(s)\right]^2$$

For convenience, let F(s) be expressed in terms of m"s and n"s.

$$F(s) = \frac{m_1'' + n_1''}{m_2'' + n_2''} = \frac{m_1'' m_2'' - n_1'' n_2''}{\left|m_2'' + n_2''\right|^2} + \frac{m_2'' n_1'' - m_1'' n_2''}{\left|m_2'' + n_2''\right|^2} \tag{11}$$

Then,

$$\left|m_2' + n_2'\right|^2 = m_2^2 + \left(2m_2 n_2 (m_2'' n_1'' - m_1'' n_2'') + n_2^2 (n_1'' - m_1'')\right) / \left|m_2'' + n_2''\right|^2 \tag{12}$$

Eq. (12) indicates that both the numerator and the denominator of eq. (10) must be multiplied by $m_2'' + n_2''$ to make EvZ(s) a ratio of polynomials.

$$EvZ(s) = \frac{(m_1 m_2 - n_1 n_2)\left|m_2'' + n_2''\right|^2}{\left|m_2' + n_2'\right|^2 \left|m_2'' + n_2''\right|^2} EvF(s) \tag{13}$$

For the even n the quadrantal zeros of EvZ(s) of rank 2n are written as

$$\prod_{i=1}^{n/2} (s^2 - s_i^2)(s^2 - \bar{s}_i^2) \tag{14}$$

Therefore eq. (14) and the numerator of eq. (13) are equated to obtain

$$EvF(s) = \frac{\prod_{i=1}^{n/2} (s^2 - s_i^2)(s^2 - \bar{s}_i^2)}{(m_1 m_2 - n_1 n_2)\left|m_2'' + n_2''\right|^2} \tag{15}$$

The roots of the polynomial $m_1 m_2 - n_1 n_2$ constitute a set of quadrantal zeros to be removed.

$$m_1 m_2 - n_1 n_2 = k(s^2 - s_1^2)(s^2 - \bar{s}_1^2) \tag{16}$$

Substitution of eq. (16) into (15) gives the desired equation.

$$EvF(s) = \frac{\prod_{i=2}^{n/2} (s^2 - s_i^2)(s^2 - \bar{s}_i^2)}{k\left|m_2'' + n_2''\right|^2} \tag{17}$$

Eq. (17) is the proof that the even part of F(s) has all the zeros of the even part of Z(s) with the first set of quadrantal zeros removed. With this proof one is justified in removing unfavorablly located sets of quadrantal zeros one by one until he obtains the F(s) belonging to the subclass of positive real functions.

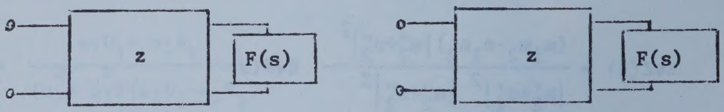
III. REALIZATION

The physical relation between two positive real functions Z(s) and F(s) is investigated next. Let us write Z(s) of eq. (9) as

$$Z(s) = \frac{m_1}{n_2}\,\frac{F(s) + n_1/m_1}{m_2/n_2 + F(s)} \tag{18a}$$

$$= \frac{m_1/m_2\ F(s) + n_1/m_2}{1 + n_2/m_2\ F(s)} \tag{18b}$$

Now, consider the two lossless two-ports terminated in F(s) and 1/F(s) respectively as shown in Fig. 2.



(a) (b)

Fig.2. Lossless two-ports with the termination of F(s) or 1/F(s)

Impedance functions of (a) and (b) of Fig. 2 are written as follows.

$$Z(s) = z_{11}\,\frac{F(s) + (z_{11}z_{22}-z_{12}z_{21})/z_{11}}{z_{22} + F(s)} \tag{19a}$$

$$Z(s) = \frac{(z_{11}z_{22}-z_{12}z_{21})F(s) + z_{11}}{1 + z_{22}F(s)} \tag{19b}$$

Noticing the correspondence between (19a) and (18a), and (19b) and (18b), and further assuming the reciprocity of lossless two-ports, one obtains the following two sets of equations.

$$z_{11} = m_1/n_2, \quad z_{22} = m_2/n_2, \quad z_{12} = \sqrt{m_1m_2 - n_1n_2}/m_2 \tag{20a}$$

Terminating function = F(s)

$$z_{11} = n_1/m_2, \quad z_{22} = n_2/m_2, \quad z_{12} = \sqrt{n_1m_2 - m_1m_2}/m_2 \tag{20b}$$

Terminating function = 1/F(s)

For convenience, let the polynomial $m_1m_2 - n_1n_2$ be denoted hereafter by $e(s^2)$. It becomes obvious that the use of mutual inductive coupling is avoided in the realization of two-port under the following cases.

Case (a) $e(s^2) = m_2^2$. Or the residue of $\sqrt{e(s^2)/n_2}$ at each pole does not exceed that of z_{22}.

Case (b) $-e(s^2) = n_2^2$. Or the residue of $\sqrt{-e(s^2)/m_2}$ at each pole does not exceed that of z_{22}.

Since both case (a) and case(b) are the constraints on the polynomial $e(s^2)$ for avoiding the use of mutual inductive coupling in realization, further examination of the polynomial is required.

$$e(s^2) = m_1m_2 - n_1n_2 = \frac{1}{(\sigma_1^2+\omega_1^2)^2} s^4 + \frac{2(\omega_1^2-\sigma_1^2)}{(\sigma_1^2+\omega_1^2)^2} s^2 + 1 \tag{21}$$

Obviously $e(s^2)$ is not a perfect square term. However it may be decomposed in the manner shown in eq. (23) so as to make each individual term correspond to either case(a) or case(b). Let eq. (9) be written as

$$Z(s) = \frac{m_1F(s) + n_1}{m_2 + n_2F(s)} = \frac{(m_{1i}F(s) + n_{1i})}{m_2 + n_2F(s)} = \sum Z_i(s) \tag{22}$$

$$(m_{1i}m_2 - n_{1i}n_2) = \sum e_i(s^2) = e(s^2) \tag{23}$$

Now case(a) and case(b) actually amount to;

$$e_1(s^2) = m_{11}m_2 - n_{11}n_2 = (Cs^2+1)^2 \tag{24}$$

$$e_2(s^2) = m_{12}m_2 - n_{12}n_2 = (Cs^2)^2 \tag{25}$$

$$e_3(s^2) = m_{13}m_2 - n_{13}n_2 = 1 \tag{26}$$

$$e_4(s^2) = m_{14}m_2 - n_{14}n_2 = -(Ds)^2 \tag{27}$$

Eq. (24) is satisfied with; $m_2 = Cs^2 + 1$, $n_2 = Ds$, $m_{11} = m_2$, $n_{11} = 0$. Thus one can write an impedance function $Z_1(s)$ as expressed by eq. (28). Since case(a) holds, eq. (20a) realizes a network such that;

$z_{11} = m_{11}/n_2$ $z_{22} = m_2/n_2$ $z_{12} = \sqrt{e_1(s^2)}/n_2 = z_{11}$ and

Terminating function = F(s).

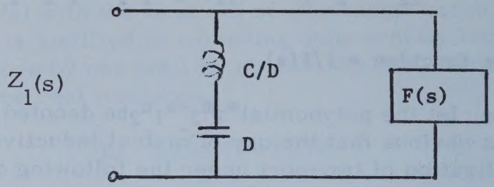


Fig. 3. Realization of $Z_1(s)$

Eq. (25) is satisfied with; $m_2 = Cs^2+1$, $n_2 = Ds$, $m_{12} = Cs^2$, $n_{12} = Cs/D$. Thus one can write an impedance function $Z_2(s)$ as expressed by **eq.(22)**. Since case(a) holds, eq. (20a) realizes a network such that; $z_{11} = m_{12}/n_2 = Cs^2/Ds$, $z_{12} = \sqrt{e_2(s^2)}/n_2 = z_{11}$ $z_{22} = (Cs^2+1)/Ds$, Terminating function = F(s).

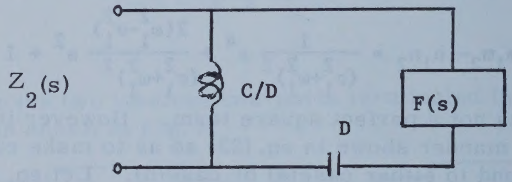


Fig. 4. Realization of $Z_2(s)$

Eq. (26) is satisfied with; $m_2 = Cs^2+1$, $n_2 = Ds$, $m_{13} = 1$, $n_{13} = Cs/D$. Thus one can write an impedance function $Z_3(s)$ as expressed by **eq.(22)**. Since case(a) holds, eq. (20a) realizes a network such that $z_{11} = m_{13}/n_2$, $z_{12} = \sqrt{e_3(s^2)}/n_2 = z_{11}$, $z_{22} = (Cs^2+1)/Ds$, Terminating function = F(s).

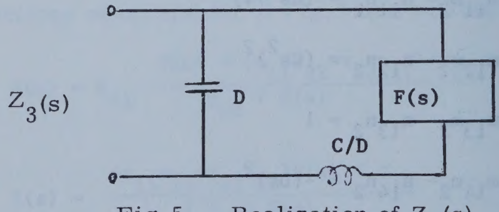


Fig. 5. Realization of $Z_3(s)$

Eq. (27) is satisfied with; $m_2 = Cs^2+1$, $n_2 = Ds$, $m_{14} = 0$, $n_{14} = Ds$, Terminating function = 1/F(s). Since case(b) holds, eq. (20b) realizes a network such that; $z_{11} = n_{14}/m_2$, $Ds/(Cs^2+1)$, $z_{12} = \sqrt{-e_4(s^2)}/m_2 = z_{11}$, $z_{22} = n_2/m_2 = z_{11}$.

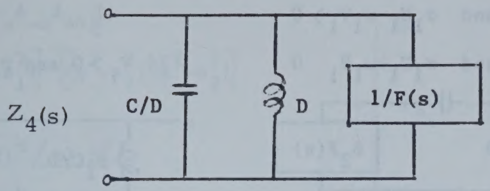


Fig. 6. Realization of $Z_4(s)$

It is obvious that if $e(s^2)$ is decomposed into three terms with multiplicative coefficients β three series connections out of four preceding networks, with impedance level raised by corresponding β; realizes $Z(s)$. It is desirable to make one of the β vanish. $Z(s)$ then is realized with series connection of two networks. Let us consider the following four cases.

$$e(s^2) = \beta_1 e_1(s^2) + \beta_2 e_2(s^2) + \beta_3 e_3(s^2) \tag{28a}$$

$$e(s^2) = \beta_1 e_1(s^2) + \beta_2 e_2(s^2) + \beta_4 e_4(s^2) \tag{29a}$$

$$e(s^2) = \beta_1 e_1(s^2) + \beta_3 e_3(s^2) + \beta_4 e_4(s^2) \tag{30a}$$

$$e(s^2) = \beta_2 e_2(s^2) + \beta_3 e_3(s^2) + \beta_4 e_4(s^2) \tag{31a}$$

Solution of eq. (28a) gives a set of equations in (28b) and so forth. From the following four sets of equations, four networks in Figs. 7, 8, 9, 10 are realized with specific condition in each network.

$$\left[\begin{array}{l} \beta_1 = (\omega_1^2 - \sigma_1^2)A \\ \beta_2 = (\sigma_1 U_1 + \omega_1 V_1)D\sqrt{A/C} \\ \beta_3 = (\sigma_1 U_1 - \omega_1 V_1)D \end{array}\right. \tag{28b}$$

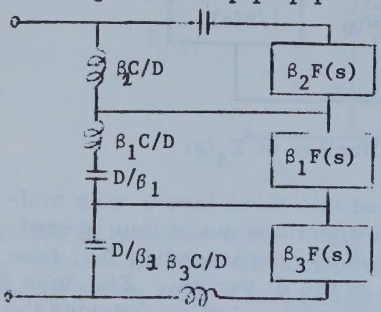
$$\left[\begin{array}{l} \beta_1 = 1 \\ \beta_2 = (4\sigma_1\omega_1 U_1 V_1)/(\omega_1 U_1 - \sigma_1 V_1)^2 \\ \beta_4 = (\sigma_1 U_1 - \omega_1 V_1)(\omega_1 U_1 - \sigma_1 V_1)^2/(\sigma_1\omega_1(\omega_1 U_1 + \sigma_1 V_1)) \end{array}\right. \tag{29b}$$

$$\left[\begin{array}{l} \beta_1 = (\omega_1 U_1 + \sigma_1 V_1)^2/(\omega_1 U_1 - \sigma_1 V_1)^2 \\ \beta_3 = 4\sigma_1\omega_1 U_1(-V_1)/(\omega_1 U_1 - \sigma_1 V_1)^2 \\ \beta_4 = (\sigma_1 U_1 + \omega_1 V_1)(\omega_1 U_1 - \sigma_1 V_1)/(\sigma_1\omega_1) \end{array}\right. \tag{30b}$$

$$\left[\begin{array}{l} \beta_2 = (\omega_1 U_1 + \sigma_1 V_1)^2/(\omega_1 U_1 - \sigma_1 V_1)^2 \\ \beta_3 = 1 \\ \beta_4 = 2(\sigma_1^2 - \omega_1^2)/(D(\sigma_1^2 + \omega_1^2))^2 \end{array}\right. \tag{31b}$$

If $V_1 > 0$ and $\sigma_1U_1-\omega_1V_1 > 0$

or $V_1 < 0$ and $\sigma_1U_1+\omega_1V_1 \quad 0$



β2C/D
β2F(s)
β1C/D
D/β1
β1F(s)
D/β3
β3C/D
β3F(s)

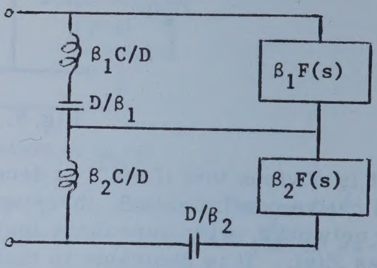
$$\beta_1 = (\omega_1^2-\sigma_1^2)A$$

$$\beta_2 = (\sigma_1U_1+\omega_1V_1)D\sqrt{A/C}$$

$$\beta_3 = (\sigma_1U_1-\omega_1V_1)D$$

Fig.7.

If $V_1 > 0$ and $\sigma_1U_1-\omega_1V_1 = 0$



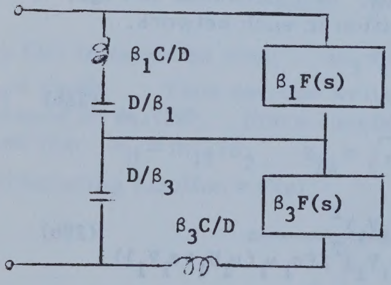
β1C/D
D/β1
β1F(s)
β2C/D
β2F(s)
D/β2

$$\beta_1 = 1$$

$$\beta_2 = 2U_1D/C(\omega_1U_1+\sigma_1V_1)$$

Fig.8.

If $V_1 < 0$ and $\sigma_1U_1+\omega_1V_1 = 0$



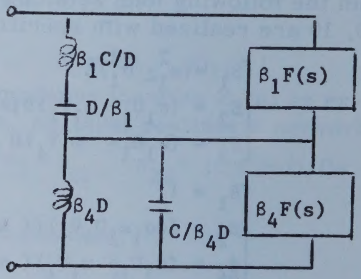
β1C/D
D/β1
β1F(s)
D/β3
β3F(s)
β3C/D

$$\beta_1 = A/C$$

$$\beta_3 = 2U_1D(-V_1)/C(\omega_1U_1+\sigma_1V_1)$$

Fig.9.

If $V_1 = 0$



β1C/D
D/β1
β1F(s)
β4D
C/β4D
β4F(s)

$$\beta_1 = 1$$

$$\beta_4 = U_1^2$$

Fig.10.

Example Synthesize the driving-point function

$$Z(s) = \frac{s^4+4s^3+11s^2+8s+6}{2s^4+7s^3+8s^2+9s+4}$$

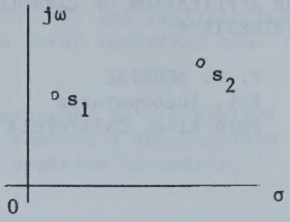
Step 1. Locate the zeros of the even part.

$$2\left[(s^4+2s^2+2)(s^4-s^2+6)\right]$$

$$= 2\left[(s^2-s_1^2)(s^2-\bar{s}_1^2)\right]\left[(s^2-s_2^2)(s^2-s_2^2)\right]$$

$$s_1 = (2)^{1/4}\,\underline{/67.5^{\circ}}$$

$$s_2 = (6)^{1/4}\,\underline{/39^{\circ}}$$



Step 2. Remove an unfavorable set of quadrantal zeros. A set of quadrantal zeros represented by s_1 is to be removed.

$A = 0.5$ $\quad \beta_1 = 0.5$

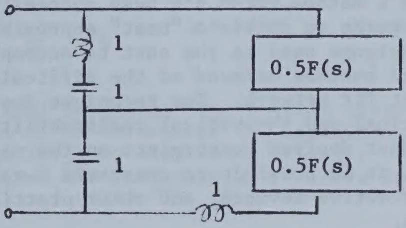
$B = 1$ $\quad \beta_3 = 0.5$

$C = 1$

$D = 0.5$

$$F(s) = \frac{(Cs^2+1)Z(s)-Bs}{(As^2+1)-DsZ(s)} = \frac{s^2+2s+3}{s^2+3s+2}$$

Step 3. Realize Z(s). Since $V_1 < 0$, and $\sigma_1 U_1 + \omega_1 V_1 = 0$, Z(s) is realized by the network of Fig.9.



REFERENCES

1. P. I. Richards: A Special Class of Functions With Positive Real Part in a Half Plane, Duke Math. J., vol. 14 No. 3, September 1947.
2. R. Bott and R. J. Duffin: Impedance Synthesis Without Use of Transformers, J. Appl. Phys., vol. 20, p. 816, August 1949.
3. O. Brune: Synthesis of a Finite Two-Terminal Network Whose Driving Point Impedance is a Prescribed Function of Frequency, J. Math. Phys., August 1931.
4. F. Miyata: A New System of Two-Terminal Synthesis, IRE Trans. on Circuit Theory, December 1955.
5. N. Balabanian: Network Synthesis, Prentice Hall, 1958.
6. D. Hazony: Zero Cancellation Synthesis Using Impedance Operators, IRE Trans. on Circuit Theory, June 1961.

"THE APPLICATION OF CARROLL'S OPTIMIZATION TECHNIQUE TO NETWORK SYNTHESIS"*

P. O. SCHEIBE
ESL, Incorporated
Palo Alto, California

E. A. HUBER
Electronic Defense Laboratories, Inc.
Sylvania Electronic Systems
Mountain View, California

ABSTRACT

The network synthesis problem is considered a minimization problem and iterative gradient techniques are used to automatically adjust the network element values to approximate a desired network response with minimum error. Contraint equations are used to insure physical realizability by use of Carroll's nonlinear programming technique. This technique transfers the minimization with constraints problem to an unconstrained minimization problem. The result is an automatic procedure which uses the circuit designer's experience in setting up the problem but requires no tedious labor on his part.

INTRODUCTION

This paper describes a method which has been successfully used to automatically adjust networks to obtain a "best" approximation to desired responses. The techniques used in the past to accomplish the adjustment have met with limited success because of the difficulty of assuring realizability of the best fit network. The technique described here circumvents both the practical and theoretical realizability problems by automatically ensuring that desired constraints on the network element values are satisfied; thus, it is possible to constrain Q values of inductors, figures of merit for active devices, and other practical network element performance criteria.

Although much of the automatic adjustment of the network involves purely mathematical procedures, certain parts of the technique involve the engineering judgment of the person using the program. These areas of judgment include specification of the desired network topology and response, the constraint functions, the error criteria, and reasonable initial element values.

CARROLL'S RESPONSE-SURFACE TECHNIQUE

Carroll's optimization technique[1] is a method which transfers the minimization with constraints problem to an unconstrained minimization problem. The Fiacco and McCormick[2,3] development of the technique as a convex programming problem is as follows: the problem is to determine a vector $\bar{C}$ to minimize $E(\bar{X})$ where $\bar{X}$ is an n component vector subject to constraints of the form of equation (1),

$$q_i(\bar{X}) \geq 0 \quad i = 1, 2, \ldots m \tag{1}$$

---

*This work was partially supported by U.S. Army Contracts DA 36-039-amc-03404 (E) and DA 28-043-amc-00379 (E). Partial support was also given by the ESL, Inc. Independent Research Program.

where $q_i(\bar{X})$ is a scalar valued function. The $q_i(\bar{X})$ are explicit functions but $E(\bar{X})$ is a composite function obtained using numerical computions.

This nonclassical minimization problem may be embedded in a classical problem of minimization without constraints by forming a new function $\phi(\bar{X};r)$ given in equation (2) where r is a real positive parameter.

$$\phi(\bar{X};r) = E(\bar{X}) + r \sum_{i=1}^{m} \frac{1}{q_i(\bar{X})} \tag{2}$$

The first term in (2) is the function to be minimized, the second term is the penalty for adding the constraints. An iterative procedure is used to minimize (2) for a strictly monotonic decreasing sequence of r values $\{r_j\}$ to obtain a sequence of points $\{\bar{C}(r_j)\}$ that respectively minimize $\phi(\bar{X};r_j)$. If we denote the feasible region as

$$R = \{\bar{X}: q_i(\bar{X}) \geq 0, \quad i = 1, 2, \ldots m\}$$

and require that $E(\bar{X})$ and $-q_i(\bar{X})$ be convex and twice continuously differentiable, it may be shown[2,4] that $\bar{C}(r_k) \to \bar{C}$ as $k \to \infty$ and

$$\lim_{k \to \infty} \phi(\bar{C}(r_k); r_k) = \inf_{\bar{X} \in R} E(\bar{X}) = \min_{\bar{X} \in R} E(\bar{X})$$

An initial point $\bar{X}°$ must be given. The initial perturbation parameter, r_o, which depends on $\bar{X}°$ is calculated and iterative gradient procedures are used to determine $\bar{C}(r_o)$ by minimizing $\phi(\bar{X};r_o)$. The perturbation parameter is reduced and the minimization repeated so that at the j^{th} stage we minimize $\phi(\bar{X};r_j)$ using $\bar{C}(r_{j-1})$ as the starting point.

APPLICATION TO THE SYNTHESIS PROBLEM

The application of this minimization technique to network synthesis is one of estimating the element values of a specified network which best meet the desired frequency specifications. The frequency specification is a plot or list of the magnitude and phase of the desired frequency function, $H(\omega)$, versus frequency. The designer is expected to define the structure of the network and describe it by a list of the type, ID number, initial value, and node connections of the elements. The values for the elements to be varied comprise the vector $\bar{X}$ with the designer's initial estimate being $\bar{X}°$. He must also set up constraint equations, $q_i(\bar{X}) \geq 0$, to insure physical realizability, ranges from 0's and any other conditions required to restrict the region in which solutions are acceptable. It should be noted that the form $q_i(\bar{X}) \geq 0$ is quite general because inequalities like $50 \leq L_5/R_1 \leq 100$ can be rewritten in the form

$$\left[(L_5/R_1) - 50\right] \left[100 - (L_5/R_1)\right] \geq 0 \quad .$$

Finally the designer must specify the measure of error, $E(\bar{X})$, between the specified frequency function, $H(\omega)$, and that obtained with the specified network structure, $\hat{H}(\bar{V}(\bar{X};\omega))$. The measure of error between the real functions $\hat{H}(\bar{V}(\bar{X};\omega))$ and $H(\omega)$ is determined by the functional f as shown in Eq. (3).

$$E(\bar{X}) = f\left(\hat{H}(\bar{V}(\bar{X};\omega)), H(\omega)\right) \tag{3}$$

The specification of the functional f causes some conceptual difficulty for the designer. One guideline in choosing f is that

$$f(\hat{H}(.), H(.)) \geq f(H(.), H(.)) = 0 \quad .$$

Choosing the "best" f, particularly when parts of the frequency band are to be given different weights, becomes somewhat an art in which experience plays a strong role.

Once the above specifications have been made, the procedure becomes automatic. The connection equations relating $\hat{H}(\bar{V}(\bar{X};\omega))$ and $\bar{X}$ are obtained automatically[5] from the description of the network graph. The admittance matrix for the network is formed using the indefinite admittance matrix[6] to internally describe devices with more than two terminals. The independent sources must be currents, $\bar{I}$, so the matrix equation

$$\bar{I} = Y(\bar{X})\bar{V} \tag{4}$$

is used to obtain the unknown voltages $\bar{V}$ as on intermediate step. Subroutines are made available to compute a variety of frequency functions, $\hat{H}(\bar{V}(\bar{X};\omega))$ using $\bar{I}$ and $\bar{V}$.

GRADIENT METHODS

The major portion of the computational effort is devoted to the minimization of $\phi(\bar{X};r_j)$. Thus, the efficiency of the technique is dependent on efficient gradient methods for determining the point $\bar{C}(r_j)$ at which the function is minimum using the initial estimate $\bar{X}^{\circ}$ or $\bar{C}(r_{j-1})$ depending on the stage of the minimization. Improved approximation to the point $\bar{C}(r_0)$ is obtained by considering the first few terms of the Taylor expansion of the function about the point $\bar{X}^{\circ}$. Consideration of the first two and three terms results in the method of steepest descent and Newton's second order gradient method, respectively. With the method of steepest descent, successive estimates of $\bar{C}$ are obtained using

$$\bar{X}^{i+1} = \bar{X}^{i} - \theta \nabla \phi (\bar{X}^{i}) \tag{5}$$

where $\nabla \phi (\bar{X})$ is the gradient vector and the scalar θ is the "distance" to move in the $-\nabla \phi$ direction. The parameter θ is determined by minimizing $\phi(\bar{X}^{i} - \theta \nabla \phi (\bar{X}^{i}))$ with respect to θ. Methods for estimating θ are discussed later. The method of steepest descent uses only first derivatives so it requires little computational effort per iteration, is exact if the surface of the function is a plane, never completely fails, but can be quite inefficient for high order surfaces. With Newton's method, successive estimates of $\bar{C}$ are obtained using

$$\bar{X}^{i+1} = \bar{X}^{i} - \theta Q^{-1} (\bar{X}^{i}) \nabla \phi (\bar{X}^{i}) \quad . \tag{6}$$

where $Q(\bar{X}^{i})$ is the matrix of second partial derivatives.

Since both first and second derivatives are required, the computational requirements per iteration grow quite rapidly with the dimension of $\bar{X}$. This method is exact for quadratic functions, converges rapidly for high order surfaces but fails on planar surfaces because Q is not positive definite. Since the initial estimate is likely to be far removed from the minimum in a region where the surface is nearly planar, the method of steepest descent makes a good initial minimization technique. As the minimum is approached, the second order terms dominate so the second order gradient method makes a good closing minimization technique.

Both of these requirements are fulfilled in a very efficient and elegant technique developed by Fletcher and Powell[8] in which successive estimates are obtained using

$$\overline{X}^{i+1} = \overline{X}^{i} - \theta H^{i} \nabla \phi (\overline{X}^{i}) \quad . \tag{7}$$

Initially, the H matrix is the identity matrix so the first iteration is identical to the method of steepest descent; that is, $H^{\circ} = I$. The changes in position and gradient for each iteration are used to modify H after each iteration. The reader is referred to Fletcher and Powell's paper[8] for the development of Eq. (7) and also a proof that if $\phi(\overline{X})$ is a quadratic function in n variables, the method will converge in n + 1 steps. The method uses only first derivatives so the computational requirements are equivalent to those for the method of steepest descent. Furthermore, as the minimum is approached and second order terms dominate, assurance is given that the H matrix will develop into the Q^{-1} matrix and the technique becomes equivalent to Newton's second order method. Thus, this method is ideally suited to the class of minimization problems being considered.

A number of iterative procedures are available for estimating θ; that is, determining the minimum of a function along a line with θ as a parameter. These range from a simple Newton-Raphson procedure[4] which minimizes the derivative to using estimates of the eigenvalues of the function.[7] Bracketing procedures which first bracket the minimum and then reduce the interval bounding the minimum are quite efficient. These include bisection and Fibonacci search[3] techniques. The bisection technique successively doubles the interval until the minimum is bracketed and then successively halves the interval to determine the minimum. The Fibonacci search procedure uses a Fibonacci sequence for bracketing and essentially divides the interval into thirds successively to determine the minimum. This technique has proved to be quite efficient.

GRADIENT VECTOR EVALUATION

A particularly convenient method of automatic evaluation of the partial derivatives in the gradient vector is given by Wengert[9]. This method is based on the chain rule,

$$\dot{u} = \frac{\partial u}{\partial x_1} \dot{x}_1 + \frac{\partial u}{\partial x_2} \dot{x}_2 + \quad \dots + \frac{\partial u}{\partial x_N} \dot{x}_N \tag{8}$$

where u is a function of the x_i, the x_i are functions of t, and the dot denotes the total derivative with respect to t. Let $\dot{x}_j = 1$ and $\dot{x}_i = 0$ for $i \neq j$; then observe that $\dot{u} = \partial u/\partial x_j$ from Eq. (8).

A set of subroutines which use elementary functions provide the analytic evaluation of the partial derivatives[10,11]. For illustrative purposes, consider the problem of numerical evaluation of the partial derivatives of

$$f = x_1 e^{x_2} \quad .$$

We construct a set of subroutines with vector-valued input and output which perform the normal elementary operations and functions. Thus, in vector notation with auxiliary vectors, z^i, define

$$z^1 = \left[z_1, \dot{z}_1\right] = \left[e^{x_2}, \dot{x}_2 e^{x_2}\right] \quad ;$$

the input to this operation is the vector, $x^2 = \left[x_2, \dot{x}_2\right]$ and the output vector is z^1. Similarly,

$$\left[f, \dot{f}\right] = \left[x_1 z_1, \dot{x}_1 z_1 + x_1 \dot{z}_1\right]$$

which requires $x^1 = \left[x_1, \dot{x}_1\right]$ and the previously calculated vector, $z^1 = \left[z_1, \dot{z}_1\right]$.

The computer programming problem is reduced to writing a set of subroutines to perform the derivative operations as illustrated above, then linking these subroutines by a set of auxiliary variables (the z^i vectors) to form the desired functions.

Application of the numerical differentiation is facilitated by observing that

$$\overline{V} = Y^{-1} \overline{I}$$

from Eq. (4) and thus

$$\dot{\overline{V}} = -Y^{-1} \dot{Y} Y^{-1} \overline{I} = -Y^{-1} \dot{Y} \overline{V} \quad . \tag{9}$$

The error function, $E(\overline{X})$, is a functional of $\hat{H}(\overline{W}(\overline{X};\omega))$ and $H(\omega)$ as noted in Eq. (3). For simplicity, but with same loss of generality, we assume that this functional may be represented as

$$E(\overline{X}) = < g\left[\hat{H}(\overline{W}(\overline{X},\omega)), H(\omega)\right], z(\omega) > \tag{10}$$

where the brackets $<,>$ represent an inner product[12], g is a scalar real-valued function which is chosen to accomplish the desired error criteria, z (the weighting function) is a scalar real-valued function of the frequency variable, ω, and $\overline{W}$ is the real vector consisting of the real and imaginary parts of $\overline{V}$,

$$\overline{W} = \left[\text{Re}V_1, \text{Re}V_2, \ldots, \text{Re}V_K, \text{Im}V_1, \text{Im}V_2, \ldots, \text{Im}V_K\right]^T \quad , \tag{11}$$

where the T denotes the transpose. The elements of the gradient of E are found by forming

$$\partial/\partial x_i \; E(\overline{X}) = \partial/\partial x_i < g\left[\hat{H}(\overline{W}(\overline{X};\omega)), H(\omega)\right], z(\omega) > \quad . \tag{12}$$

The partial derivative in Eq. (12) is found to be

$$\partial/\partial x_i \; g\left[\hat{H}(\overline{W}(\overline{X};\omega)), H(\omega)\right] = \partial g/\partial \hat{H} \sum_j \partial w_j/\partial x_i \; \partial \hat{H}/\partial w_j \tag{13}$$

through use of the chain rule.

Each of the partial derivatives, $\partial w_j/\partial x_i$, in Eq. (13) must now be obtained from the network equations. Since each of the element values, x_i, is taken to be a real number,

$$\partial w_j/\partial x_i = \begin{cases} \partial/\partial x_i \; \text{Re}\, V_j = \text{Re}\, \partial V_j/\partial x_i & j = 1, 2, \ldots, K \\ \partial/\partial x_i \; \text{Im}\, V_{j-K} = \text{Im}\, \partial V_{j-K}/\partial x_i & j = K+1, \ldots, 2K \end{cases} \tag{14}$$

Evaluation of Eq. (9) gives

$$\dot{\overline{V}}^i = \left[\partial V_1/\partial x_i, \partial V_2/\partial x_1, \ldots, \partial V_K/\partial x_i\right]^T = -Y^{-1} \dot{Y}\Big|_{\dot{x}_j = \delta_{ij}} \overline{V} \tag{15}$$

where $\dot{Y}$ is evaluated with all $\dot{x}_j = 0$ except $\dot{x}_i = 1$. Define the vector, $\overline{F}$, by

$$\bar{F}^{g} = \left[f_1{}^{g}, f_2{}^{g}, \ldots, f_{2K}{}^{g}\right]^T = \left[(\mathrm{Re}\ \dot{\bar{V}}^{g})^T \,\middle|\, (\mathrm{Im}\ \dot{\bar{V}}^{g})^T\right]^T \qquad (16)$$

then,

$$f_j{}^{i} = \partial w_j / \partial x_i \qquad (17)$$

by comparing Eqs. (14), (15) and (16). Substitute Eq. (17) into Eq. (13)

$$\partial/\partial x_i\, g\left[\hat{H}(\bar{W}(\bar{X};\omega)), H(\omega)\right] = \partial g/\partial H \ \langle \bar{F}^{i}, \nabla_{\bar{W}}\hat{H} \rangle \qquad (18)$$

where $\langle , \rangle$ denotes the finite dimensional euclidean inner product over the enclosed vectors. Combining Eq. (18) and (12) one obtains

$$\nabla_{\bar{X}}E(\bar{X}) = \begin{bmatrix} <\partial g/\partial H \ \langle F^1, \nabla_W H \rangle, z(\omega)> \\ \vdots \\ \vdots \\ <\partial g/\partial H \ \langle F^n, \nabla_W H \rangle, z(\omega)> \end{bmatrix} \qquad (19)$$

Use of Wengert's method in Eq. (15) thus allows one to form the vector $\bar{F}i$ and finally with the aid of Eq. (19), obtain $\nabla_{\bar{X}}E(\bar{X})$. Finally, with Eq. (2) the gradient of ϕ $(\bar{X};r)$ may be written

$$\nabla_{\bar{X}}\phi(\bar{X};r) = \nabla_{\bar{X}}E(\bar{X}) - r \sum_{i=1}^{m} \left[1/q_i(\bar{X})\right]^2 \nabla_{\bar{X}}q_i(\bar{X}) \quad , \qquad (20)$$

where Eq. (19) is used in the evaluation of $\nabla_{\bar{X}}E(\bar{X})$ and Wengert's method is used to obtain the $\nabla_{\bar{X}}q_i(\bar{X})$. If g is a function of several $\hat{H}_j$ the method outlined above may be simply extended by further use of the chain rule.

INITIALIZATION AND TERMINATION STRATEGIES

Precise mathematical statements can be made about initialization and termination procedures which guarantee a correct solution if the conditions previously stated are satisfied. In practice, the concern over obtaining this solution efficiently and the tendency to ignore these conditions, particularly the strictly convex error function criterion, make these procedures more of an art than a science.

The efficiency of the optimization technique is dependent on the choice of the initial point $\bar{X}°$. If the designer is required to expend much effort in estimating $\bar{X}°$, the main advantage of the approach is lost. When the feasible region is convex, the initial estimate is not critical because it can be improved by minimizing with $r = 0$ until a boundary is intersected or the true minimum is reached. If a boundary is intersected, the position at the previous step will make a good initial point for starting the constrained minimization. Another possibility which has worked well on similar types of problems is to use a Monte Carlo technique to obtain an initial estimate.

Two problems associated with the sequence $\{r_i\}$ are the choice of an initial value, r_o, and a constant divisor, d, such that $r_{i+1} = r_i/d$. If the conditions previously stated are satisfied, the only requirements are that $r_o > 0$ and $d > 1$. However, the total number of required iterations is highly dependent on r_o and slightly dependent on d. The method presently used for estimating r_o is to minimize $|\nabla\phi(\bar{X}°;r)|^2$ with respect to r as suggested by Fiacco and McCormick.[3] The total number of required iterations do not appear to be highly dependent on the factor d which sets the convergence rate of $\{r_j\}$. Values from 10 to 100 have been used for d with the observation that large changes in r make large changes in the surface and require many iterations to determine $\bar{C}(r_j)$. However, the total number of r_j's need for a given accuracy in C decreases so the total number of iterations and total computer time is not highly sensitive to the parameter d.

With iterative numerical solutions which do not produce an exact result after a finite number of steps it is necessary to define an appropriate termination test which controls the final set of calculations. If the input data is "noisy" the theories of hypothesis testing may be applied to obtain a statistically reasonable termination point. From statistical theory[13], an appropriate termination test is to form the positive number

$$\xi^j = \left[E(\bar{X}^{j-1}) - E(\bar{X}^j)\right] / E(\bar{X}^j)$$

then terminate the iterations when ξ^j is less than a preassigned number, ε. The value of ε is usually chosen to be commensurate with the precision of the numerical calculations (e.g., $\varepsilon = 10^{-6}$ if a precision of 8 digits is carried in the calculations). After the termination test is satisfied, a final change in the $\bar{X}_i$ are made utilizing the fact that near the local minimum, the error surface is quadratic.[2] Since the error function, $E(\bar{X})$, is frequently not a convex function, multiple local minima may be found. Suitable choice of the initial point is the best procedure to prevent convergence to a local minimum which is not the global minimum.

EXAMPLE

The fifteen node, two stage, band-pass amplifier circuit shown in Figure 1 is used as an illustrative example. The variable parameters are the values of the eight elements encircled in Figure 1; the initial estimates are listed below

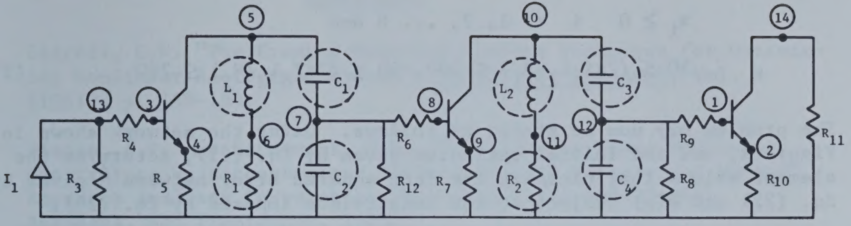
$$\begin{array}{lll} x_1 = R1 = .35 \text{ ohms} & \quad & x_5 = C3 = 83 \text{ pf} \\ x_2 = R2 = .40 \text{ ohms} & & x_6 = C4 = 16 \text{ pf} \\ x_3 = C1 = 91 \text{ pf} & & x_7 = L1 = .075\ \mu\text{h} \\ x_4 = C2 = 16 \text{ pf} & & x_8 = L2 = .083\ \mu\text{h} \end{array} \qquad (21)$$

The other element values are nominal for this type of circuit and will not be listed for lack of space. The equivalent circuit used for the transistors is also shown in Figure 1 and the values used are nominal for a 2N918.

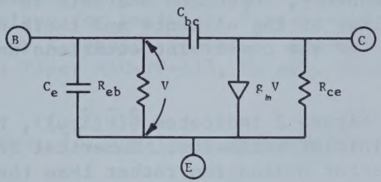
The frequency function of interest is

$$\hat{H}(\bar{V}(\bar{X};\omega)) = \left| V_{14}/V_{13} \quad (\bar{X};\omega) \right|^2 \qquad (22)$$

where V_{14}, V_{13} are the voltages at nodes 14 and 13 respectively in Figure 1. The specified frequency function, $H(\omega)$, is the solid trapezodial curve shown in Figure 2. The error function is made the mean square error between $\hat{H}$ and H with equal weights for the 21 frequency points shown in Figure 2. The constraint equations are


5
10
14
13
3
4
R4
R3
R5
I1
L1
R1
6
7
C1
C2
8
9
R6
R12
R7
L2
R2
11
12
C3
C4
1
2
R9
R8
R10
R11

TWO STAGE BAND-PASS AMPLIFIER


Cbc
B
C
Ce
Reb
V
gmV
Rce
E

TRANSISTOR EQUIVALENT CIRCUIT

FIGURE 1

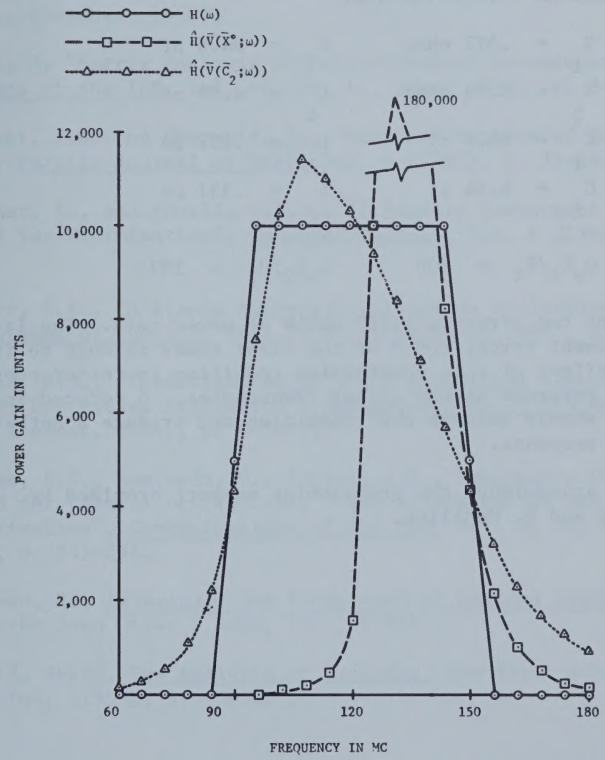


FIGURE 2

$x_i \geq 0 \quad i = 1, 2, \ldots 8$ and

$$50 \leq (2\pi f_o L_1/R_1) \leq 200, \quad 50 \leq (2\pi f_o L_2/R_2) \leq 200 \qquad (23)$$

The problem may now be stated as follows. Using the network shown in Figure 1, and the initial estimates given by Eq. (21), determine the element values that minimize the mean squared error between $\hat{H}$ given by Eq. (22) and $H(\omega)$ subject to the constraints imposed by Eq. (23).

Before describing the problem to the program, the standard normalization with respect to frequency, impedance and gain is performed. The circuit is described by a list of the elements and their node connections, a subroutine is written for the constraint equations and a list of $H(\omega)$ is provided.

The dashed plot in Figure 2 indicates $\hat{H}(\overline{V}(\overline{X};\omega))$, that is, the response obtained with the initial estimates. Numerical derivatives were used for the gradient vector evaluation rather than the method described above because it has not been fully implemented. With 44 iterations, that is 44 different direction calculations, including 2 values of r the response, $\hat{H}(\overline{V}(\overline{C};\omega))$, is indicated by the dotted plot in Figure 2. The element values at this point are:

R_1 = .572 ohms C_3 = 84.5 pf

R_2 = 2.44 ohms C_4 = 4.59 pf

C_1 = 90.8 pf L_1 = .152 μh

C_2 = 6.56 pf L_2 = .337 μh

$\omega_o L_1/R_1 = 200$ $\omega_o L_2/R_2 = 103$

The resultant rms error is 1,710 units of power gain. The large in-band gain requirement forces the Q of the first tuned circuit to its maximum value; the effect of this constaining condition is rather poor fit of the desired response at the higher frequencies. A reduced in-band gain requirement should release the constraint and produce a better fit to the desired response.

Finally, we acknowledge the programming support provided by P. Grossberg and C. McMillan.

REFERENCES

1. Carroll, C.W. "The Created-Response-Surface Technique for Optimizing Nonlinear Restrained System", Operations Research, Vol. 9 (1961), p. 169-184.

2. Fiacco, A.V. and McCormick, G.P., "Programming Under Nonlinear Constraints by Unconstrained Minimization: A Primal-Dual Method", Research Analysis Corporation Technical Paper RAC-TP-96; Bethesda, Md; (September, 1963).

3. Fiacco, A.V. and McCormick, G.P., "The Sequential Unconstrained Minimization Technique for Nonlinear Programming. Algorithim II, Optimum Gradients by Fibonacci Search." Research Analysis Corporation Technical Paper RAC-TP-123, McLean, Virginia; (June, 1964).

4. Saaty, T.L. and Bram, J., Nonlinear Mathematics, McGraw-Hill, 1964, p. 141-144, 56-70.

5. Huber, E.A. and Grossberg, P.I., "A Computer Program for the Frequency Analysis of Linear Electrical Networks", Sylvania Electronic Defense Laboratories Technical Memorandum EDL-M752, Mountain View, Calif., (October, 1964).

6. Shekel, J. "Matrix Analysis of Multi-Terminal Transducers", Proceedings of the IRE, Vol. 42, No. 5., (May, 1954) p. 840-847.

7. Crockett, J.B. and Chernoff, H., "Gradient Methods of Maximization", Pacific Journal of Math, Vol. 5 (1955), p. 33-50.

8. Fletcher, R., and Powell, M.J.D., "A Rapidly Convergent Descent Method for Minimization", Computer Journal, Vol. 6 (1963), p. 163-168.

9. Wengert, R.E., "A Simple Automatic Derivative Evaluation", Communications of the ACM, Vol. 7, No. 8 (August, 1964), p. 463-464.

10. Wilkins, R.D., "Investigation of a New Analytical Method for Numerical Derivative Evaluation", Communications of the ACM, Vol. 7, No. 8 (August, 1964), p. 465-471.

11. Bellman, R.E., Kagiwada, H., Kalaba, R.E., "Wengert's Numerical Method for Partial Derivatives, Orbit Determination and Quasilinearization", Communications of the ACM, Vol. 8, No. 4 (April, 1965), p. 231-232.

12. Friedman, B., Principles and Techniques of Applied Mathematics, New York: John Wiley & Sons, Inc. (1956).

13. Scheffé, Henry, The Analysis of Variance, New York: John Wiley & Sons, Inc. (1959), p. 25-51.

SOME RESULTS CONCERNING LOW-PASS LADDER NETWORKS

F. F. KUO
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

ABSTRACT

In this paper some ramifications of the use of convergents in ladder network analysis are discussed. There are two main results. The first deals with an iterative formula for the input impedance of a low-pass ladder network. The second is concerned with synthesis of a given low-pass ladder in inverse order. Finally, a formula on the sensitivity of an input impedance with respect to a load admittance is derived.

INTRODUCTION

In a previous note [1] some results concerning the use of convergents [2],[3],[4] in the synthesis of mth root ladder networks were discussed. We will explore in this paper some further ramifications of convergents in the analysis of low-pass ladder networks.

Let the nth convergent be given by:

$$\frac{P_n}{Q_n} = b_o + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \ldots + \frac{a_n|}{|b_n}. \tag{1}$$

The following iterative formulas apply

$$\frac{P_n}{Q_n} = \frac{b_n P_{n-1} + a_n P_{n-2}}{b_n Q_{n-1} + a_n Q_{n-2}} \qquad n \geq 2 \tag{2}$$

and

$$P_n Q_{n-1} - Q_n P_{n-1} = (-1)^{n+1} a_1 a_2 \ldots a_n \tag{3}$$

Applying these formulas to the ladder network in Fig. 1, the input impedance is

$$Z_n(s) = \frac{P_n(s)}{Q_n(s)} = \frac{y_n(s)P_{n-1}(s) + P_{n-2}(s)}{y_n(s)Q_{n-1}(s) + Q_{n-2}(s)} \tag{4}$$

where $Z_{n-1} = P_{n-1}/Q_{n-1}$ and $Z_{n-2} = P_{n-2}/Q_{n-2}$ are the input impedances with (n-1) and (n-2) arms respectively. Note that P_i and Q_i are, in general, rational. If we represent $Z_n(s)$ as a ratio of polynomials,

$$Z_n(s) = \frac{N(s)}{D(s)}$$

then, except in special cases,

$$\begin{aligned} N(s) &\neq y_n(s)P_{n-1}(s) + P_{n-2}(s) \\ D(s) &\neq y_n(s)Q_{n-1}(s) + Q_{n-2}(s). \end{aligned} \tag{5}$$

APPLICATION TO LOW-PASS LADDER FILTERS

A more convenient form of (4) is obtained when we restrict our attention to low-pass LC or RC networks with all zeros of transmission at infinity. Then, if the nth element is a shunt arm,

$$Z_n(s) = \frac{N_n(s)}{D_n(s)} = \frac{y_n N_{n-1}(s) + N_{n-2}(s)}{y_n D_{n-1}(s) + D_{n-2}(s)} \tag{6}$$

where N_i and D_i represent polynomials, N_{n-1}/D_{n-1} is the short-circuit input impedance and N_{n-2}/D_{n-2} is the open circuit input impedance. A similar expression is obtained when the nth element is a series arm. It is also easily shown that

$$N_n D_{n-1} - N_{n-1} D_n = (-1)^{n+1} \tag{7}$$

If we regard the nth element as a load impedance, then (6) describes a relation between input and load impedances. Note that (6) involves only input impedance terms and can be contrasted to the standard formula relating Z_{in} and Z_L

$$Z_{in}(s) = z_{11} + \frac{z_{12}z_{21}}{z_{22} + Z_L} \tag{8}$$

which involves all four open-circuit parameters. Using the equivalent circuit in Fig. 2, (8) can be quite simply derived from (6).

Next let us consider a rather unusual result concerning doubly terminated low-pass filters. We restrict our attention to the type of ladder networks with all zeros of transmission at $s = \infty$. Turning to Fig. 3a, let us denote the short-circuit input impedance by $Z_s = N_s/D_s$ and the open-circuit input impedance by $Z_o = N_o/D_o$. We then define a new impedance function

$$Z_x(s) = \frac{N_s(s) - D_s(s)}{N_o(s) - D_o(s)}. \tag{9}$$

We can show that $Z_x(s)$ is not only realizable but can be synthesized as a ladder whose elements are obtained in inverse order as those of the original filter. The only exception is that the source resistance of the original filter and the load resistance of the Z_x filter differ by one ohm. For the network in Fig. 3a, the Z_x network is given in Fig. 3b. The proof of this concept is quite easily obtained. Using the iterative relationship given in (6) we obtain both (N_s-D_s) and (N_o-D_o) as functions of (N_1-D_1). Then we expand $Z_x(N_1-D_1)$ in a continued fraction to obtain the elements. Consider the example of Fig. 4a.

$$Z_s = \frac{N_s}{D_s} = \frac{120s^3 + 52s^2 + 49s + 8}{24s^3 + 8s^2 + 9s + 1} \tag{10}$$

$$Z_o = \frac{N_o}{D_o} = \frac{40s^3 + 4s^2 + 15s + 1}{8s^3 + 3s} \tag{11}$$

Using Z_x defined in (9), we obtain

$$Z_x(s) = \frac{96s^3 + 44s^2 + 40s + 7}{32s^3 + 4s^2 + 12s + 1}$$

$$= 3 + \frac{1|}{|s} + \frac{1|}{|4s} + \frac{1|}{|2s} + \frac{1|}{|4}. \qquad (12)$$

The realization of $Z_x(s)$ is given in Fig. 4b.

It should be noted that the concept just discussed is applicable to RL and RC ladder networks, also. The only restrictions are: (1) all zeros of transmission should be at $s = \infty$, (2) the source impedance should be resistive and have a value exceeding one ohm.

SENSITIVITY

For the low-pass ladder whose input impedance is given by $Z_n(s)$ in (6), we can easily determine the sensitivity of the input impedance with respect to the load immittance y_n. Using the standard Bode-Mason definition, the sensitivity function is

$$S_{yn} = \frac{d \ln Z_n}{d \ln y_n}$$

$$= \frac{y_n N_{n-1}(s)}{N_n(s)} - \frac{y_n D_{n-1}(s)}{D_n(s)} \qquad (13)$$

which, according to (7) can be simplified to

$$S_{yn} = \frac{y_n(-1)^n}{N_n(s)D_n(s)}. \qquad (14)$$

Expanding S_{yn} in (13) in partial fractions, we obtain

$$S_{yn} = \sum_{j=0}^{m} \frac{S_{yn}^{\xi_j}}{(s-\xi_j)} - \sum_{i=0}^{k} \frac{S_{yn}^{\gamma_i}}{(s-\gamma_i)} + K_o \qquad (15)$$

where γ_i and ξ_j are the poles and zeros of $Z_n(s)$, $S_{y_n}^{\gamma_i}$ and $S_{y_n}^{\xi_j}$ represent respectively the pole and zero sensitivities and K_o is a real constant. Thus, as Ur [5] and Mikulski [6] have shown, we can express input impedance sensitivity as a weighted sum of pole and zero sensitivities.

REFERENCES

[1] F. F. Kuo, "The Use of Convergents in the Synthesis of Mth Root Ladder Networks" to be published in Transactions of the IEEE on Circuit Theory, September, 1965.

[2] A Ya Khinchin, "Continued Fractions," University of Chicago Press, Chicago, Illinois, 1964.

[3] A. N. Khovanskii, "The Application of Continued Fractions and Their Generalizations to Problems in Approximation Theory," (Translated by Peter Wynn), P. Noordhoff, N. V. Groningen, the Netherlands, 1963, Chapters I and IV.

[4] O. Perron, "Die Lehre von den Kettenbruchen," B. G. Teubner, Leipzig, 1913, Chapters I and IV.

[5] H. Ur, "Geometrical Properties of Root Loci," M. S. Thesis, Purdue University, 1956.

[6] J. J. Mikulski, "Correlation Between Classical and Pole-Zero Sensitivity," Doctoral Dissertation, University of Illinois, 1959.

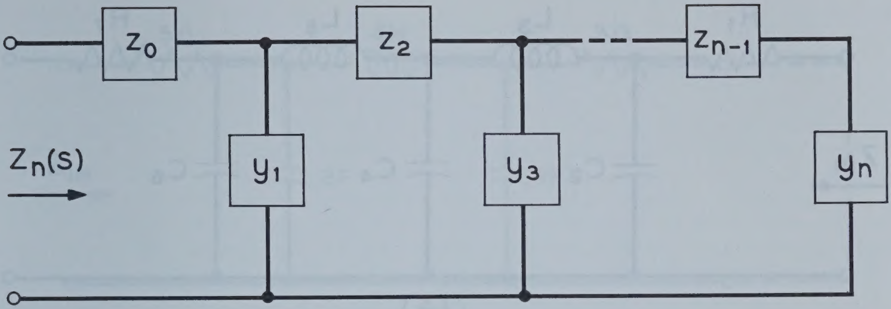

Z_0
Z_2
Z_{n-1}
$Z_n(S)$
y_1
y_3
y_n

FIG. 1

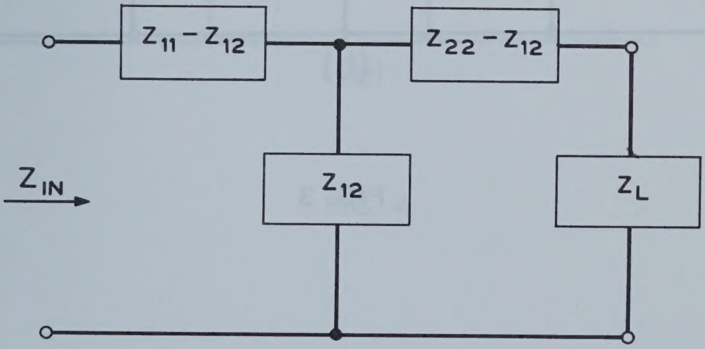

$Z_{11} - Z_{12}$
$Z_{22} - Z_{12}$
Z_{IN}
Z_{12}
Z_L

FIG. 2

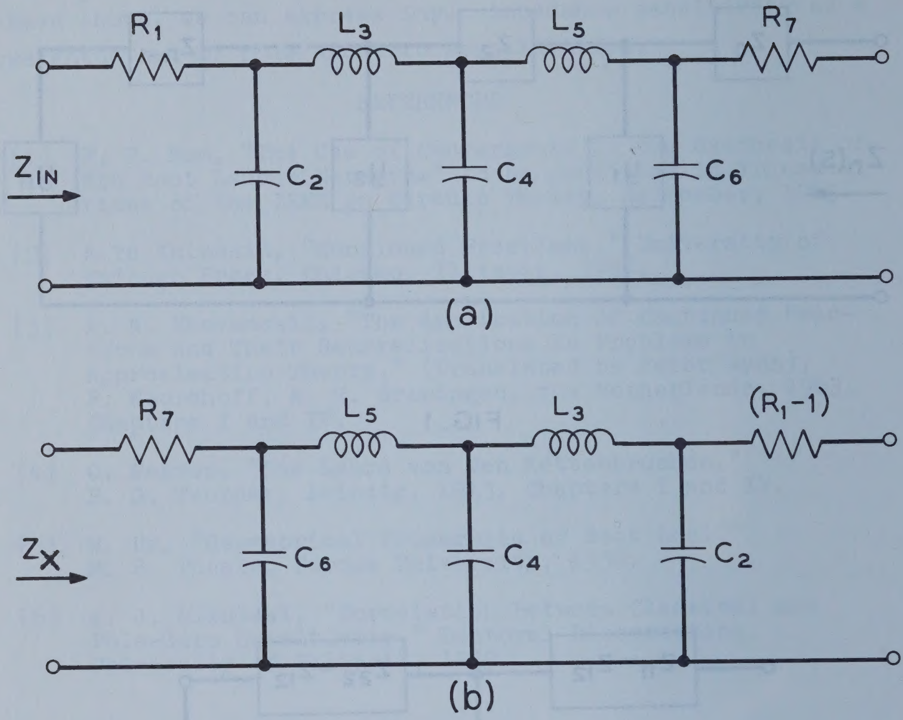

R1
L3
L5
R7
ZIN
C2
C4
C6
(a)
R7
L5
L3
(R1-1)
ZX
C6
C4
C2
(b)

FIG. 3

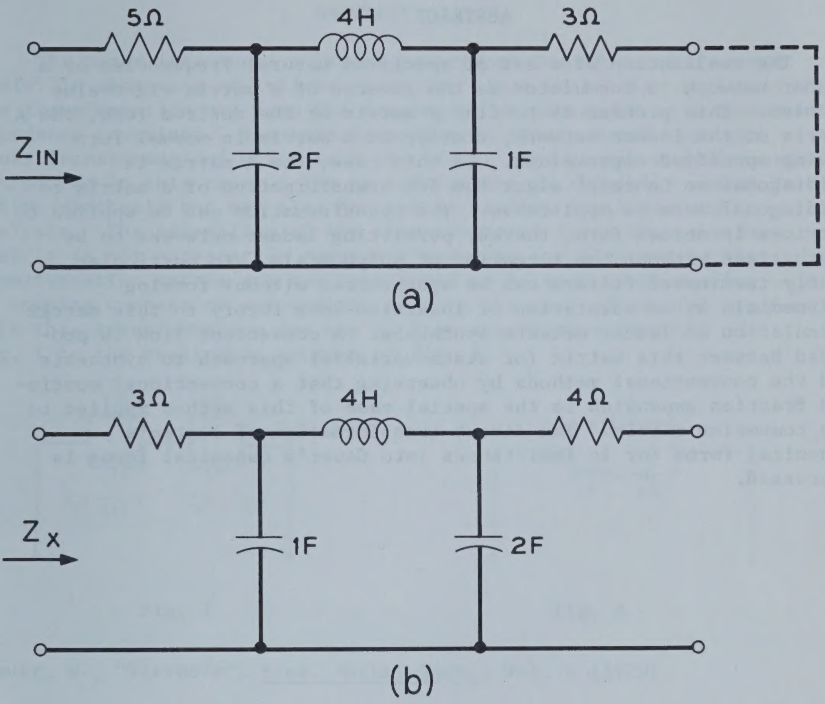

5Ω
4H
3Ω
Z_IN
2F
1F
(a)
3Ω
4H
4Ω
Z_X
1F
2F
(b)

FIG. 4

LADDER NETWORK SYNTHESIS BY SIMILARITY TRANSFORMATION

THOMAS G. MARSHALL, JR.
Radio Corporation of America
P. O. Box 800
Princeton, New Jersey

ABSTRACT

The realization of a set of specified natural frequencies by a ladder network is formulated as the inverse of a matrix eigenvalue problem. This problem is to find a matrix of the desired form, the A matrix of the ladder network, similar to a matrix in normal form having specified eigenvalues. In this case, the A matrix is tridiagonal so Lanczos' algorithm for transformation of a matrix to tridiagonal form is applicable. The transformation can be applied to matrices in normal form, thereby permitting ladder networks to be synthesized without the formation of polynomials. In particular, doubly terminated filters can be synthesized without forming polynomials by an adaptation of insertion-loss theory to this matrix formulation of ladder network synthesis. A convenient link is provided between this matrix (or state-variable) approach to synthesis and the conventional methods by observing that a conventional continued fraction expansion is the special case of this method applied to the companion matrix. The direct transformation of Foster's canonical forms for lc immittances into Cauer's canonical forms is discussed.

A GENERAL TRANSFORMATION PROCEDURE IN CIRCUIT THEORY

MARCO A. MURRAY-LASSO
Bell Telephone Laboratories
Whippany, New Jersey

ABSTRACT

The paper introduces a Linear Functional Transformation method useful in obtaining equivalent circuits. As a particular application the transformation may keep the driving point impedance or a transfer impedance invariant. The transformation is more general than the Cauer Transformation[1] which is a particular case of the transformation introduced in this paper. The transformation is specially useful in active synthesis but may also be used in passive synthesis and analysis. The generality of the transformation stems from the fact that it is not necessarily congruent and that the elements of the transformation matrix are not restricted to be real constants but may be complex numbers or polynomials of s in the case of passive synthesis. With this transformation one may go from the "Foster Circuit" of Fig. 1 to the realizable network of Fig. 2.

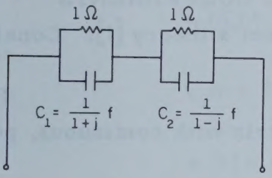


Fig. 1

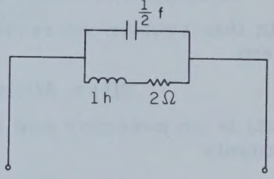


Fig. 2

---

[1] Cauer, W., "Vierpole", Elek. Nachr. Tech., Vol. 6 (1929).

SENSITIVITY ANALYSIS WITH RANDOM INCREMENTS

A. PAPOULIS
Polytechnic Institute of Brooklyn
Brooklyn, New York

ABSTRACT

It is known that a linear systems $\dot{x}(t) = A(t)\,x(t)$ with periodic coefficients $A(t + T) = A(t)$ is stable if the eigenvalues $\lambda_i(0)$ of the impulse response matrix $\Phi(t)$ evaluated at $t = T$, lie inside the unit circle. Thus, to determine the stability of the system one must find $\Phi(T)$, usually by a numerical integration. Due to various errors the computed matrix is $\Phi(T) + \varepsilon\overline{\Phi}$ with eigenvalues $\lambda_i(\varepsilon)$ and the problem arises how to determine $\overline{\Phi}$ and its effect on λ_i. In the first part of this paper the more general problem of evaluating the displacement of the singularities of a network function due to incremental variation in its elements is considered. As a special case, $\overline{\Phi}$ is related to $\lambda_i(\varepsilon)$. In the second part, the accumulation of round-off and truncation errors in the numerical solution of a non-linear system $\dot{x}(t) = f(x, t)$ is estimated and the results are used to evaluate the perturbation $\overline{\Phi}$.

LINEAR SYSTEMS WITH PERIODIC COEFFICIENTS

In this section we review briefly Floquet's theory [1]. Consider the system

$$\dot{x}(t) = A(t)\,x(t) \tag{1}$$

where x(t) is an n-vector and A(t) an n x n matrix with continuous, periodic elements

$$A(t + T) = A(t) \tag{2}$$

We denote by $\Phi(t)$ its impulse response matrix (transition matrix, or weighting pattern), i. e., a set of n solutions of (1) with initial conditions

$$\Phi(0) = I \quad \text{(unit matrix)} \tag{3}$$

Clearly, any solution x(t) of (1) is given by

$$x(t) = \Phi(t)\,x(0) \tag{4}$$

Since

$$\dot{\Phi}(t + T) = A(t + T)\,\Phi(t + T) = A(t)\,\Phi(t + T) \tag{5}$$

we conclude that $\Phi(t + T)$ is also a solution matrix. Its initial condition is $\Phi(T)$, hence, [see (4)]

$$\Phi(t + T) = \Phi(t)\,\Phi(T) \tag{6}$$

From the above follows that

$$\Phi(t + mT) = \Phi(t)\,\Phi^m(T) \tag{7}$$

for any integer m. Therefore, $\Phi(t)$ is bounded if $\Phi^m(T)$ is bounded for $m > 0$ (one-sided stability). One can show (using, e. g., z transform) that this is the case if the eigenvalues λ_i of $\Phi(T)$

$$|\Phi(T) - \lambda_i I| = 0 \tag{8}$$

lie inside the unit circle.

The above stability condition can also be established by showing that (1) has n solutions of the form

$$\psi_i(t) = \lambda_i^{t/T}\ \pi_i(t) \qquad \pi_i(t + T) = \pi_i(t) \tag{9}$$

where λ_i are the n eigenvalues of (8), which we assume simple (the multiple case can be handled similarly). Clearly, if $|\lambda_i| < 1$ then $\psi_i(t)$ is bounded and if this is true, then any solution of (1) is also bounded because the vectors $\psi_i(t)$ are obviously independent.

To prove (9) we denote by e_i the eigenvectors of $\Phi(T)$

$$\Phi(T)\ e_i = \lambda_i e_i \tag{10}$$

With $\psi_i(t)$ the solution of (1) with initial condition e_i

$$\psi_i(0) = e_i$$

we have [see (4)]

$$\psi_i(t) = \Phi(t)\ \psi_i(0) = \Phi(t) e_i \tag{11}$$

But $\psi_i(t + T)$ is also a solution with initial conditions [see (10) and (11)]

$$\psi_i(T) = \Phi(T)\ \psi_i(0) = \lambda_i \psi_i(0)$$

hence,

$$\psi_i(t + T) = \lambda_i \psi_i(t) \tag{12}$$

With

$$\pi_i(t) = \psi_i(t) \lambda_i^{-(t/T)}$$

we thus have [see (12)]

$$\pi_i(t + T) = \psi_i(t + T) \lambda_i^{-(1 + t/T)} = \psi_i(t) \lambda_i^{-(t/T)} = \pi_i(t)$$

and (9) follows.

Thus, to determine the stability of (1) one must first compute the matrix $\Phi(T)$, then find the roots λ_i of (8). Due to local round-off and truncation errors in a numerical integration of (1), the computed matrix is $\Phi(T) + \varepsilon \overline{\Phi}$. The accumulated error $\varepsilon \overline{\Phi}$ causes an error in λ_i and the problem arises how to determine $\overline{\Phi}$ and its effect on λ_i. Motivated by the above, we are considering in the following, two related problems with more general applications.

THE SENSITIVITY PROBLEM IN NETWORK THEORY

A basic problem in network synthesis and control theory [2] is the determination of the displacement Δp of the singularities of the system function, resulting from variations in its elements. This problem is usually analysed by considering the effect of the variation of each element on Δp and summing the resulting displacements. [3, 4] This approach is complicated and it might even lead to erroneous orders of magnitude. In the following we present a method for determining Δp directly in terms of the variation of all network elements.

The Characteristic Equation

As a preparation, we shall relate the roots of the polynomial

$$D_o(p) \quad \text{degree } n \tag{13}$$

to the roots of the perturbed polynomial

$$D(p) = D_o(p) + \varepsilon D_1(p) + \ldots + \varepsilon^k D_k(p) \tag{14}$$

We denote by p_o a root of $D_o(p)$ and by $p(\varepsilon)$ the "neighboring" root of $D(p)$ in the sense

$$\lim p(\varepsilon) = p_o \qquad \varepsilon \rightarrow 0 \tag{15}$$

Simple root. Suppose that $D_o'(p_o) \neq 0$. It is not difficult to show that, in this case, $p(\varepsilon)$ can be written as a series

$$p(\varepsilon) = p_o + p_1\varepsilon + \ldots \tag{16}$$

converging for sufficiently small ε. If $D_1(p_o) \neq 0$, then

$$p_1 = -\frac{D_1(p_o)}{D'(p_o)}$$

If the first non-zero term in the expansion (14) of $D(p_o)$ is $D_r(p_o)$, then

$$p(\varepsilon) = p_o - \frac{\varepsilon^r D_r(p_o)}{D'(p_o)} + O(\varepsilon^{r+1}) \tag{17}$$

Multiple roots. If p_o is of multiplicity h and $D_1(p_o) \neq 0$, then

$$\frac{D_o^{(h)}(p_o)}{h!}\left[p(\varepsilon) - p_o\right]^h = -\varepsilon D_1(p_o) + O(\varepsilon^2)$$

From the above it follows that there are h simple roots of $D(p)$ "near" p_o and each can be written as a series (Puisoux's expansion [5])

$$p^i(\varepsilon) = p_o + p^i_{1/h}\sqrt[h]{\varepsilon} + p^i_{2/h}\sqrt[h]{\varepsilon^2} + \ldots \tag{18}$$

in powers of $\sqrt[h]{\varepsilon}$. The first term is given by

$$p^i_{1/h} = \sqrt[h]{-\frac{\varepsilon h!\, D_1(p_o)}{D_o^{(h)}(p_o)}} \qquad i = 1, \ldots, h \tag{19}$$

Thus, if $D_1(p_o) \neq 0$, then the roots of $D(p)$ lie on the circumference of a circle with center p_o, for small ε. If $D_1(p_o) = 0$ and the first non-zero term is $D_r(p_o)$, we can no longer proceed as in the case of a simple root. To determine the first non-zero term in the expansion of $p(\varepsilon) - p_o$ we must use the expansion of every $D_s(p)$ for $s \leq r$ and solve the solve the resulting equation. For small ε it is no longer true that the roots $p^i(\varepsilon)$ lie on a circle, in fact, they might all be real.

Fundamental comment. If it is known that the roots of $D(p)$ are real for any real ε, then the expansion (18) contains only integral powers of ε

$$p^i(\varepsilon) = p_o + p^i_1\varepsilon + p^i_2\varepsilon^2 + \ldots \qquad i = 1, \ldots, h \tag{20}$$

This follows easily from the fact that $p(\varepsilon)$ and $p(-\varepsilon)$ is real.

Singular case. Suppose that the degree m of $D(p)$ is greater than the degree n of $D_o(p)$. In this case, $D(p)$ has m roots, of which n are "near" the roots of $D_o(p)$ and can be found as above. The remaining $\nu = m-n$ roots are large for small ε and can be evaluated as follows. Dividing (14) by p^m we obtain with $s = 1/p$

$$B(s) = s^{\nu} B_o(s) + \varepsilon B_1(s) + \dots + \varepsilon^k B_k(s) \tag{21}$$

where

$$B_o(s) = s^n D_o(1/s) \quad B_i(s) = D_i(1/s) \quad i = 1, \dots, k$$

The unperturbed term $s^{\nu} B_o(s)$ in (21) has a root $s = 0$ of multiplicity ν generating ν roots for $B(s)$ near $s = 0$. Assuming $B_1(0) \neq 0$, we obtain as in (19)

$$s^i(\varepsilon) \simeq \sqrt[\nu]{-\frac{\varepsilon B_1(0)}{B_o(0)}} \qquad i = 1, \dots, \nu \tag{22}$$

hence

$$p^i(\varepsilon) \quad \sqrt[\nu]{-\frac{B_o(0)}{\varepsilon B_1(0)}} \tag{23}$$

i.e., $p^i(\varepsilon)$ tends to infinity as $\varepsilon^{-1/\nu}$.

Perturbation of the Characteristic Roots of a Network Matrix

We shall assume that the given network has only two kinds of elements (the three element case can be handled similarly): Its singularities are the roots of the determinant

$$|A + pB| \tag{24}$$

where A and B are symmetrical nxn matrices (e.g., the G and C matrices in nodal analysis.) Our problem is to relate the roots of (24) to the roots of the perturbed determinant

$$|(A + pB) + \varepsilon(\overline{A} + p\overline{B})| \tag{25}$$

where $\varepsilon\overline{A}$ and $\varepsilon\overline{B}$ are the matrices of the perturbation of the network elements.

We denote by p_o a root of (24) and by x_o the corresponding eigenvector

$$|A + p_oB| = 0 \qquad (A + p_oB)x_o = 0 \tag{26}$$

To the root p_o there corresponds a root $p(\varepsilon)$ of (25)

$$\left|\left[A + p(\varepsilon)B\right] + \varepsilon\left[\overline{A} + p(\varepsilon)\overline{B}\right]\right| = 0 \tag{27}$$

neighboring p_o in the sense (15) with $x(\varepsilon)$ the corresponding eigenvector

$$\left(\left[A + p(\varepsilon)B\right] + \varepsilon\left[\overline{A} + p(\varepsilon)\overline{B}\right]\right) x(\varepsilon) = 0 \tag{28}$$

From the symmetry of the given matrices it follows that p_o and $p(\varepsilon)$ are real, hence, we can assume that x_o and $x(\varepsilon)$ are also real.

Simple root. If p_o is a simple root, then $p(\varepsilon)$ can be written as a series [see (16)]

$$p(\varepsilon) = p_o + p_1\varepsilon + p_2\varepsilon^2 + \dots \tag{29}$$

As we see from (28), the components of the vector $x(\varepsilon)$ are ratios of polynomials in ε, hence, $x(\varepsilon)$ can also be expanded

$$x(\varepsilon) = x(0) + x_1\varepsilon + x_2\varepsilon^2 + \dots \tag{30}$$

But $\lim p(\varepsilon) = p_o$, $\lim x(\varepsilon) = x(0)$ for $\varepsilon \longrightarrow 0$ therefore $x(0)$ is an eigenvector of $A + p_oB$ and since its eigenvectors are unique within a factor, we can assume $x(0) = x_o$. Inserting (30) and (29) into (28) and collecting equal powers of ε, we obtain (26) and

$$(p_1 B + \overline{A} + p_o\overline{B})\, x_o + (A + p_oB)\, x_1 = 0 \tag{31}$$

$$(p_2B + p_1\overline{B})x_o + (p_1B + \overline{A} + p_o\overline{B})x_1 + (A + p_oB)x_2 = 0 \tag{32}$$

Using these expansions one can evaluate successively the coefficients in (29) and (30). With $\{x, y\}$ the inner product of two vectors, we conclude from (26) and the symmetry of A and B that

$$\left\{(A + p_oB)x_1,\ x_o\right\} = \left\{x_1,\ (A + p_oB)x_o\right\} = 0 \tag{33}$$

Forming the inner product of (31) with x_o and using (33) we obtain $\left\{(p_1B + \overline{A} + p_o\overline{B})x_o, x_o\right\} = 0$ hence

$$p_1 = \frac{\left\{(\overline{A} + p_o\overline{B})x_o, x_o\right\}}{\left\{Bx_o, x_o\right\}} \tag{34}$$

With p_1 so determined, we solve (31) for x_1; forming the ineer product of (32) with x_o we find

$$p_2 = -\frac{\left\{(p_1B + \overline{A} + p_o\overline{B})x_1, x_o\right\} + \left\{p_1\overline{B}x_o, x_o\right\}}{\left\{Bx_o, x_o\right\}} \tag{35}$$

and so on.

Returning to (8) we observe the if the impulse matrix $\Phi(T)$ is perturbed by $\varepsilon\overline{\Phi}$, then the eigenvalue $\lambda_i(\varepsilon)$ of $\Phi(T) + \varepsilon\overline{\Phi}$ can be determined as above with

$$A = \Phi(T) \quad B = I \quad \overline{A} = \Phi \quad \overline{B} = 0$$

The resulting first order correction is [see (34) and (10)]

$$\lambda_i(\varepsilon) = \lambda_i - \varepsilon\frac{\left\{\overline{\Phi}e_i, e_i\right\}}{||e_i||}$$

where $||e_i|| = \{e_i, e_i\}$ is the length of e_i.

Multiple roots. Suppose that p_o is a root of (24) of multiplicity h. Since the roots of (25) are real, we conclude from an earlier remark that $p(\varepsilon)$, hence also $x(\varepsilon)$, are given by series with integral powers of ε as in (29) and (30). However, unlike the case of a simple root, $x(0)$ depends on the perturbing matrices $\overline{A}$ and $\overline{B}$, and our problem is to find it. As we know, to the root p_o we can associate h linearly independent eigenvectors $z_1, \dots, z_h$

$$(A + p_o B) z_i = 0 \qquad z_i z_k = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases} \tag{36}$$

Since $x(0)$ is also an eigenvector of $A + p_o B$, it must be of the form

$$x(0) = \alpha_1 z_1 + \dots + \alpha_h z_h \tag{37}$$

To determine the coefficients α_i we observe that $\{(A + p_o B)x_1, z_i\} = 0$ as in (33). From this and (31) it follows that

$$\{(p_1 B + \overline{A} + p_o \overline{B})\, x(0),\ z_i\} = 0 \quad i = 1, \dots, h \tag{38}$$

Inserting $x(0)$ as given by (37) and using (36) we obtain

$$\sum_{k=1}^{h} \{(p_1 B + \overline{A} + p_o \overline{B}) z_k, z_i\} \alpha_k = 0 \quad i = p, \dots, h \tag{39}$$

This is a homogeneous system of h equations is α_k and it has a nontrivial solution if the determinant of its coefficients equals zero

$$\begin{vmatrix} \{(p_1 B + \overline{A} + p_o \overline{B}) z_1, z_1\}, & \dots, & \{(p_1 B + \overline{A} + p_o \overline{B}) z_h, z_1\} \\ \{(p_1 B + \overline{A} + p_o \overline{B}) z_1, z_h\}, & \dots, & \{(p_1 B + \overline{A} + p_o \overline{B}) z_h, z_h\} \end{vmatrix} = 0 \tag{40}$$

The above is an equation of degree h for the variable p_1 and it has h roots p_1^i. These root give the first term in the expansions

$$p^i(\varepsilon) = p_o + p_1^i \varepsilon + p_2^i \varepsilon^2 + \dots \quad i = 1, \dots, h \tag{41}$$

To each p_1^i there corresponds an eigenvector $\alpha^i = (\alpha_1^i, \dots, \alpha_n^i)$ of (40), determining the term

$$x^i(0) = \alpha_1^i z_1 + \dots + \alpha_h^i z_h$$

in the expansions

$$x^i(t) = x^i(0) + x_1^i \varepsilon + x_2^i \varepsilon^2 + \dots \tag{42}$$

of the eigenvector $x^i(\varepsilon)$ corresponding to $p^i(\varepsilon)$.

If all the roots of (40) are distinct, then the perturbed equation (25) has h distinct roots near p_o. If not, then within first order powers of ε (25) has also multiple roots. With $x^i(0)$ and p_1^i so determined, the evaluation of the remaining coefficients of (41) and (42) can be carried out as in the non-degenerate case.

Singular case. Suppose that the degree m of the perturbed determinant (25) exceeds the degree n of (24). In this case, (25) has n roots near the roots of (24) and $\nu = m-n$ large roots. To find them we divide (24) and (25) by p^m and set $p = 1/s$.

$$|s A + B| = 0 \tag{43}$$

$$|(s A + B) + \varepsilon(s \overline{A} + \overline{B})| = 0 \tag{44}$$

Equation (43) has a zero at $s = 0$ of multiplicity ν. The neighboring roots $s^i(\varepsilon)$ of (44) are of the form

$$s^i(\varepsilon) = s_1^i \varepsilon + s_2^i \varepsilon^2 + \dots \qquad i = 1, \dots, \nu$$

and can be found as in (41). The "large" roots of (25) are given by $p^i(\varepsilon) = 1/s^i(\varepsilon)$ and tend to infinity as $1/\varepsilon$.

ACCUMULATION OF ERRORS IN THE NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS

Consider the system

$$\dot{x}(t) = f(x, t) \qquad x(0) = a \tag{45}$$

of n non-linear equations where x, f, and a are n-vectors. In a numerical integration of (45) a number of errors are introduced. As a result, the computed functions y(t) differ from x(t) by a total error

$$e(t) = y(t) - x(t) \tag{46}$$

and our objective is to determine this error [6, 7]. The main points of this discussion can be explained for $n = 1$ the conclusions are easy to generalize to any n. The usual integration methods are essentially of the following form: A number of points t_i, usually equidistant, are selected in the (0, t) interval

$$t_o = 0 \qquad t_m = t \qquad T = t_{i+1} - t_i$$

and $y(t_{i+1})$ is evaluated from $y(t_i)$, possibly other preceeding values $y(t_{i-r})$, and the function f(x, t). The computed $y(t_{i+1})$ differs from the solution of the given equation, passing through $y(t_i)$, by a "truncation" error ε_i^t depending on the method of integration, and a "round-off" error ε_i^r depending on the available storage capacity of the computer. The total local error at the i-th step is thus given by

$$\varepsilon_i = \varepsilon_i^t + \varepsilon_i^r$$

This error does not, of course, equal the difference between $y(t_{i+1})$ and the exact solution $x(t_{i+1})$ of (1) with initial condition $x(0) = a$.

With

$$b(t) = \sum_{i=1}^{m} \varepsilon_i \delta(t - t_i) \tag{47}$$

it is easy to see that y(t) is the exact solution of

$$\dot{y}(t) = f(y, t) + b(t) \qquad y(0) = a \tag{48}$$

Hence,

$$\dot{e}(t) = f(y, t) - f(y-e, t) + b(t) \qquad e(0) = 0 \tag{49}$$

We shall assume that the errors are sufficiently small so that

$$f(y, t) - f(y - e, t) = f_x(y, t)e \tag{50}$$

With this approximation, (49) yields

$$\dot{e}(t) = f_x(y, t)\, e(t) + b(t) \qquad e(0) = 0 \tag{51}$$

This is a linear equation in e(t) with time-varying coefficient the known function $f_x(y, t)$ where y(t) is known from the integration of (1). Thus, e(t) is the output of an initially inert time-varying linear system with input a sequence of impulses b(t). With $h(t;\tau)$ its impulse response,

$$e(t) = \int_0^t h(t;\tau)\, b(\tau) d\tau = \sum_{i=1}^{m} \varepsilon_i h(t;t_i) \tag{52}$$

We shall replace b(t) by the continuous function

$$n(t) = \sum_{k=1}^{m} \varepsilon_k \frac{\sin \pi(t/T - k)}{\pi(t - kT)} \tag{53}$$

If T is small, the effect of this approximation on e(t) is negligible. It is easy to see that $h(t;\tau) = h(t;0)/h(\tau;0)$, hence,

$$e(t) = h(t;0) \int_0^t \frac{n(\tau)}{h(\tau;0)} d\tau \tag{54}$$

The function h(t;0) can be evaluated by solving (45) again with initial condition a + δ. With $y_\delta(t)$ the resulting solution, e(t) is given by

$$\delta h(t;0) = y_\delta(t) - y(t) \qquad y_\delta(0) = a + \delta \tag{55}$$

The constant δ should be small enough so as to remain in the region of linearity. This can be tested numerically by computing $y_{2\delta}(t)$. The condition is

$$y_{2\delta}(t) - y(t) \simeq 2\left[y_\delta(t) - y(t)\right]$$

The truncation error is generally independent of i or T

$$\varepsilon_i^t = \varepsilon^t$$

The round-off error ε_i^r is a function of T and tends to zero, usually faster than T

$$\varepsilon_i^r = f(T)$$

From (53) we see that

$$n(kT) = \frac{\varepsilon_k}{T} = \frac{f(T) + \varepsilon^t}{T}$$

This shows that a decrease of T does not necessarily reduce the total error e(t). There is an optimum spacing depending on f(T). If

$$f(T) = \alpha T^2 \quad \text{than} \quad \alpha T_{opt}^2 = \varepsilon^t$$

We shall assume that the local errors ε_i are independent random variables with mean η_i and variance σ_i^2

$$\eta_i = E\{\varepsilon_i\} \qquad \sigma_i = E\{(\varepsilon_i - \eta_i)^2\}$$

As we see from (51) the total error e(t) is a Markoff process. With

$$\eta(\tau) = \sum_{k=1}^{n} \eta_k \frac{\sin \pi(t/T-k)}{\pi(t-kT)} \qquad \sigma^2(\tau) = \sum_{k=1}^{n} \sigma_k^2 \frac{\sin \pi(t/T-k)}{\pi(t-kT)}$$

we easily find

$$E\left\{e(t)\right\} = h(t;\, 0) \int_0^t \frac{n(\tau)}{h(\tau;\, 0)}\, d\tau$$

$$\sigma_e^2 = h^2(t;\, 0) \int_0^t \frac{\sigma^2(\tau)}{h^2(\tau;\, 0)}\, dr$$

The preceding equations hold also for a system of n equations. In this case, x, y, f, e, ε_i, b, a are n-vectors and f_x in (51) is an nxn matrix with elements

$$\frac{\partial^{i} f_k(x_1, \ldots, x_m, t)}{\partial x_i}$$

where f_k is the k-th component of f. The impulse response h(t;τ) is an nxn matrix satisfying

$$h(t;\, \tau) = h(t;\, 0)\, h^{-1}(\tau;\, 0)$$

and (52) reads

$$e(t) = \int_0^t h(t;\, 0) h^{-1}(\tau;\, 0)\, b(\tau) d\tau \tag{56}$$

To find e(t) it suffices therefore to know h(t; 0). This matrix can be computed by solving (45) n times. Wtih $(a_1, \ldots, a_n)$ the components of x(0) = a we determine numerically a set of n solutions forming the matrix Y_δ with initial conditions

$$Y_\delta(0) = \begin{pmatrix} a_1+\delta, & a_2, & \ldots, & a_n \\ \cdots & \cdots & \cdots & \cdots \\ a_1, & a_2, & \ldots, & a\ \ +\delta \end{pmatrix}$$

The impulse matrix h(t;0) is then given by

$$\delta h(t;0) = Y_\delta(t) - Y_o(t)$$

We return to (1). The numerical integration yields a matrix Y(t) differing from Φ(t) by the error matrix E(t) = Y(t) - Φ(t) satisfying

$$\dot{E}(t) = A(t)\, E(t) + B(t) \tag{57}$$

where B(t) is due to the local errors [see (47)]. The impulse response matrix of (57) is the unknown matrix Φ(t). Within first order errors Φ(t) can be replaced by the computed matrix Y(t), hence

$$E(t) = \int_0^t \Phi(t)\, \Phi^{-1}(\tau)\, B(\tau) dr \simeq \int_0^t Y(t)\, Y^{-1}(\tau) B(\tau)\, d\tau$$

The error $E(T) = \varepsilon \overline{\Phi}$ can then be used to determine the perturbations of the eigenvalues of λ_i.

REFERENCES

[1] Lefschetz, S. "Differential Equations: Geometric Theory," Interscience Publishers, New York,

[2] Tomovic, R., "Sensitivity Analysis of Dynamic Systems," McGraw-Hill, 1964.

[3] Desoer, C. A., Network Design by First-Order Predistortion Technique, IRE, PGCT, September 1957.

[4] Papoulis, A., Displacement of the Zeros of the Impedance Z(p) Due to Incremental Variations in the Network Elements, Proc. IRE, January 1955.

[5] Rellich, F., "Perturbation Theory of Eigenvalue Problems," Inst. of Math. Sciences, N. Y. Univ., 1953.

[6] Rademacher, H., "On the Accumulation of Errors in Process of Integration of High-Speed Calculating Machines," Proc. of the Symposium on Large Scale Digital Calculating Machinery (1948). Harvard Computational Lab. Series, Vol. 16.

[7] Papoulis, A., "On the Accumulation of Errors in the Numerical Solution of Differential Equation," Journal of Applied Ph., February 1952.

SOME APPROACHES TO THE REDUCTION OF CONTROL SYSTEM SENSITIVITY

P. KOKOTOVIĆ, S. BINGULAC and J. MEDANIĆ
Institute "Mihailo Pupin"
Belgrade, Yugoslavia

INTRODUCTION

The long-neglected problem of synthesis of systems insensitive to parameter variations has recently received increasing attention[1,2,15]. In applying the classical procedures of frequency response, root loci and parametric plane, attempts to take into account some sensitivity performance criterion are made[3,4]. Parallel to the revival of the old idea of feedback loops with infinite gains, some new principles, for example, systems with changing structures, have been proposed[5]. Some work has been initiated on the synthesis of adaptive systems from the viewpoint of minimizing the sensitivity to parameter variations[6,7]. From the current investigations it appears that special attention should be assigned to the following:

- formulation of performance criteria so as to incorporate the demands of minimizing sensitivity;
- development of an applicable design procedure satisfying or minimizing the chosen criteria;
- investigation of typical structures from the viewpoint of both sensitivity and the practical requirement of a simple and realizable structure as a result of the synthesis.

To summarize, this field of automatic control is still in its embryonic stage, and before a general, but sufficiently applicable and practical procedure is developed, additional experience should be accumulated.

This paper discusses some experience in solving the synthesis problem in the time domain where performance criteria are defined with the help of time domain sensitivity functions. In the first section the sensitivity points are employed to find an appropriate structure of the controller. In the second section, fixed structure is assumed and the procedure of parametric minimization, employing a combined error-sensitivity criteria, is demonstrated. Lastly, in the third section we proceed from the minimization procedure accomplished in the course of system design to adaptive structures where the parameters are adjusted in the course of operation. The sensitivity functions of the non-linear adaptive systems are compared with the ones obtained from the previous linear systems.

SENSITIVITY COMPARISON OF TWO TYPICAL LINEAR STRUCTURES

If the plant to be controlled is such that it permits a certain freedom in the selection of the controller structure the designer has to determine which structure is favourable with respect to the sensitivity demands. The sensitivity functions

$$u_i(t) = \frac{\partial y(t)}{\partial \ln p_i} \qquad (1)$$

where y(t) is the system response and p_i plant parameters, as well as the disposition of the sensitivity points, may be quite helpful. The sensitivity points are defined as those points in the system up to which, starting from the input of the system, the transmittance is equal to the logarithmic sensitivity function

$$S^T_{p_i} = \frac{\partial \ln T}{\partial \ln p_i} \qquad (2)$$

T = T(s) is the system transmittance.

To be specific, suppose that a typical structure is employed in connection with the plant p_{123} in Fig. 1. As a result of the sensitivity points method, the time domain sensitivity functions $u_i(t)$ with respect to gain coefficients of the corresponding branches of the plant may be measured at the sensitivity points N_i (Fig. 1), if the output signal of the system is applied to the input of the second example of the system or its model - sensitivity model. A diagram of the sensitivity functions is given in Fig. 2 for a particular system and the case of a step input. Evidently, the parameter p_3 has a dominant effect on the overshoot and response time. For example, if p_3 receives an increment of 20% the overshoot will increase from 6% to about 12%. The sensitivity to p_2 is much weaker and to p_1 even more. Of the utmost importance is the fact that for a great majority of systems with the structure as in Fig. 1 such a gradation exists between the sensitivities with respect to the particular branches and this may be considered a general property of this structure.

Suppose the variations of p_3 are most probable while the variations of p_1 are in this sense least critical. In that case, such a relation between sensitivities is unfavourable. In fact, just the opposite would be much more desirable. This suggests selecting a structure in which the branches exchange position, Fig. 3. It is easily ascertained that this is exactly the structure in Fig. 4 representing again the real plant. This can be shown by inverting the direction of all the arrows on the graph, and the graph in Fig. 3 may be used as a sensitivity model of the graph in Fig. 4. If the transmittances $c^*_i h^*_i$ in the obtained system in Fig. 4 are so selected that the response of the system is similar to the response of the system in Fig. 1, the sensitivity functions for the same plant are given in Fig. 5. Clearly the sensitivity to parameter p_3 is diminished for several times. In retaliation we are compelled to accept an increase in sensitivity with respect to p_1 and p_2. It should be mentioned that in the described example the transmittances of the feedback branches were of the simplest form

$$c_i h_i(s) = \frac{c_i s}{c_{ii} s + 1}$$

and the elements of the plant were of the first or second order. A series of such experiments confirmed the well known practical rule, that the more feedback loops we put around some particular element, the less is the system sensitive to its changes. Unfortunately, in practice, the input into the last, most critical part of the plant is usually inaccessible to the control action and the structure of Fig. 4 is rarely employed.

COMBINED ERROR-SENSITIVITY CRITERIA

In the more realistic case, the choice of the system structure is restricted and a simple controller is demanded. Thus the problem of synthesis reduces into parametric optimization of some performance criteria. Analytic and computer methods for the minimization of error criteria, for example integral square error, are well known[8,9,10]. If the demand for minimization or restriction of sensitivity is to be satisfied in the design, the possibility of employed a combined performance criterion

$$J = J_1 + \mu J_2 = \int_0^\infty e^2 dt + \mu \int_0^\infty u_i^2 dt \qquad (3)$$

has been investigated, where u_i is the sensitivity with respect to a variable plant parameter. If one of the indices J_1 or J_2 are constrained to some allowed value, the problem reduces to constrained optimization and may be solved by the method of Lagrange multipliers. The procedure is similar to minimization with the constraint of control action demonstrated by Newton et al[8]. Indeed, if the system and its sensitivity model are examined together, the sensitivity function is a full-fledged coordinate of the system. From a different point of view the same performance criterion is minimized in the work of Šiljak and Burzio[11] with the condition of a sufficient relative stability of the system. Both approaches, although simple in principle, should be examined critically from the practical point of view. Analytical minimization is complicated by the fact that the order of sensitivity function U(s) is twice the order of the observed system. In Ref. 11 the difficulties are reduced by the application of the parameter plan technique and tables of the integral square error.

The possibility of automatic minimization of the criterion on computers with the application of the sensitivity points method is discussed in this paper.

Employing the gradient procedure parameter c_i is adjusted in accordance with the equation

$$\frac{\partial J}{\partial c_i} = -h \int_0^\infty y v_i dt - h\mu \int_0^\infty u_i w_i dt \qquad (4)$$

where h is a positive constant, $v_i = \partial y/\partial c_i$ the sensitivity function of the first order and $w_{ij} = \partial u_i/\partial c_j$ sensitivity function of the second order.

The procedure for analog computer minimization of the criterion J_1 is completely described in some earlier papers[10,12] and the basic advantage is the possibility of obtaining all the components of the gradient simultaneously by using the system and an additional sensitivity model. If the minimization of the combined criteria is to be practically acceptable, it is necessary that the sensitivity functions of the second order may be obtained in a similar and simple manner. It can be shown that for the structure in Fig. 1 this is possible with the implementation of one additional model[13]. In other words, the equipment necessary for the application of the procedure consists of the model of the system, two additional sensitivity models, and the corresponding elements for the adjustment of parameters.

A convenient way of performing the procedure is to minimize the criterion J for a distinct value of μ and then vary μ from 0 to ∞, obtaining in such way the "μ trajectory." The optimum values of the parameters should be selected from the μ trajectory on which different values of J_1 and J_2 are calibrated. Depending on the relative importance of the two indices the optimum point may be nearer the minimum of J_1 or the minimum of J_2. An example of the μ-trajectory for a linear system of the fifth degree is given in Fig. 6. The sensitivity functions and the response of the system corresponding to three points on the μ-trajectory are given in Fig. 7a, b, c. Evidently, the insistence on the minimization of sensitivity would in this case lead to a degradation of the system output. Thus, the investigation of the system performance along the μ-trajectory fully answers the question concerning what may be achieved by parametric optimization with a fixed structure in regard to the minimization of sensitivity. Unfortunately, this and a number of other examples indicate that these possibilities in systems with fixed structures are not too common.

SENSITIVITY OF ADAPTIVE SYSTEMS

Up to now the sensitivity of linear systems has been analyzed and the results applied to the design and optimization of ordinary linear control systems. Now, we propose a criterion for adaptive systems used to control slowly varying plants. The systems discussed belong to the model - reference adaptive type with the application of a sensitivity model. The manner of control to be applied will depend on the degree of an a priori knowledge about the plant. In our case the plant is not totally unknown; in fact, a great deal is known about it. The global structure of the plant plus controller is assumed to be known, as well as the nominal or expected values of plant parameters and the range of possible plant characteristics variations, although the actual manner of the variations is not assumed or known in advance. The change in plant characteristics will be manifested through variations of its parameters or even changes in the structure of the plant, all as a consequence of a changing environment.

A basic question that always arises when some semidefined object is to be controlled is its identification, since this will greatly influence the manner of a control. Intuitively one feels that since the ignorance is not complete the identification does not have to be complete either if the a priori information is utilized.

The a priori knowledge about the nominal values and the range of parameter variations, together with the proposed structure of the controller permits the avoidance of explicit identification of the actual state of the system. It is sufficient to know which way to proceed to get near the optimum and the range of convergence of the applied algorithm. The algorithm used in slow adaptation and optimization is the gradient method leading to a type of adaptive control systems called the model-reference adaptive control systems with the implementation of a sensitivity model[7,14]. A feasible way to validate the applicability of an algorithm for a particular adaptive system is to simulate the whole system on an analog computer. Some elegant structures of adaptive systems have been obtained and one of them is shown in Fig. 8. Such time-varying systems will be non-linear due to the existing multiplication in the system. The analysis of such systems is incredibly difficult and a general approach has yet to be perceived.

Nevertheless, the idea of employing the information that may be obtained at the sensitivity points should not be deserted because of the mere fact that we are not able to express mathematically the phenomena taking place in such a system. The main question is to adapt the system and the secondary question to find a mathematical justification for the procedure.

What we propose is to ignore the time-varying characteristics due to changes in plant parameters and regard it as stationary, and extend the methodology to time-varying systems. The system obtained in this manner may be treated as a system to which a number of non-linear loops has been added with the intention of making its characteristics insensitive to the variations in plant parameters. The selection of the filters in the adaptive loops is a question by itself and demands separate attention. In a preliminary study we have found it convenient to choose a controller filter of the type

$$A_i(s) = a_{1i} + \frac{a_{2i}}{a_{3i}s + 1} \tag{5}$$

where a_{ij} $(j = 1,2,3)$ are chosen in such a way to give satisfactory adaptation.

The possibilities and shortcomings of such non-linear and non-stationary systems have yet to be envisioned. The stability analysis will hardly be possible by the existing methods, and for now we have had to be satisfied with the fact that the proposed system is much more insensitive to plant parameter variations than without the adaptive loops.

The very statement above points out the criterion used in estimating the quality of the adaptive loops. The sensitivity functions with respect to the variable plant parameters have been analyzed and the result was that the sensitivity functions (or a sensitivity index based on them) are much smaller with the adaptive loops than without them. With the system in the optimum state, the sensitivity to the plant parameter variations was lessened in a great degree in comparison to the sensitivity without the adaptive loops, by the sole presence of the loops (Fig. 9). The reduction in the sensitivity was even greater if the system was not in its optimum state, the adaptive loops being effective (Fig. 10). Incidentally, this sensitivity index was also used to select the optimum parameters of the filter in (5), the optimum values being those minimizing the sensitivity to plant parameter variations.

Thus, the sensitivity approach to the adaptation of time-varying plants gives possibilities in the synthesis of model-reference adaptive systems. The approach may well become useful as a procedure for setting up intentionally non-linear feedback loops with the purpose of desensitizing the system to plant parameter variations.

REFERENCES

1. Proceedings of IFAC/ETAN Symposium on Sensitivity Analysis, Dubrovnik, Sept. 1964, edited by Lj. Radanovic, Pergamon Press 1965.
2. Kokotovic, P. and Rutman, R., "The Sensitivity of Automatic Control Systems", Avtomatika i Telemehanika, No. 4, April 1965.
3. Horowitz, I. M., "Synthesis of Feedback Control Systems", Academic Press, New York, London, 1963.
4. Siljak, D. D., "Analysis and Synthesis of Feedback Control Systems in the Parameter Plane, Pts. I, II, III", IEEE Trans. Pt. II (Application and Industry), Vol. 83, No. 75, Nov. 1964, pp. 449-473.
5. Emelianov, C. B. and Bermant, M. A., "On the Obtaining the High Performance Control of Time Varying Plants", Proc. Academy of Sc. USSR, Tom 145, No. 4, 1962.
6. Bingulac, S. P., "Minimum Sensitivity Adaptive Systems", Proc. of IFAC/ETAN Symposium on Sensitivity Analysis, Dubrovnik 1964.
7. Kokotovic, P. V., Medanic, J. V. and Vuskovic, M. I., Bingulac, S. P., "Sensitivity Method in the Experimental Design of Adaptive Control Systems", will be presented at the 3rd IFAC Congress, London, July 1966.
8. Newton, G. C., Gould, L. A. and Kaiser, J. F., "Analytical Design of Feedback Control", John Wiley, 1957.
9. Siljak, D. D., Ph.D. Thesis, University of Belgrade, 1963.
10. Bingulac, S. and Kokotovic, P., "Automatic Optimization of Linear Control System on an Analog Computer", Proc. of the Inter. Assoc. for Analog Computation, Vol. VII, No. 1, January 1965 (also presented at the 8th ETAN Conference, 1963).
11. Siljak, D. D. and Burzio, A., "Minimization of Sensitivity with Stability Constraints in Linear Control Systems", will be presented at the Third Allerton Conference, Urbana, Illinois, Oct. 20-22, 1965.
12. Kokotovic, P. "Sensitivity Points Method in the Analysis and Optimization of Linear Control Systems", Avtomatika i Telemehanika, No. 12, Dec. 1964, Moscow.
13. Bingulac, S. P., "Simultaneous Generation of Sensitivity Functions of Second Order, IBK Report 1965.

14. Osborn, P. V., Whitaker, H. P. and Kezer, A., "New Developments in the Design of Model Reference Adaptive Control System", IAS Paper No. 61-39, New York, January 1961.

15. Cruz, J. B. and Perkins, W. R., "A New Approach to the Sensitivity Problem in Multivariable Feedback System Design", Trans. IEEE, AC, July 1964.

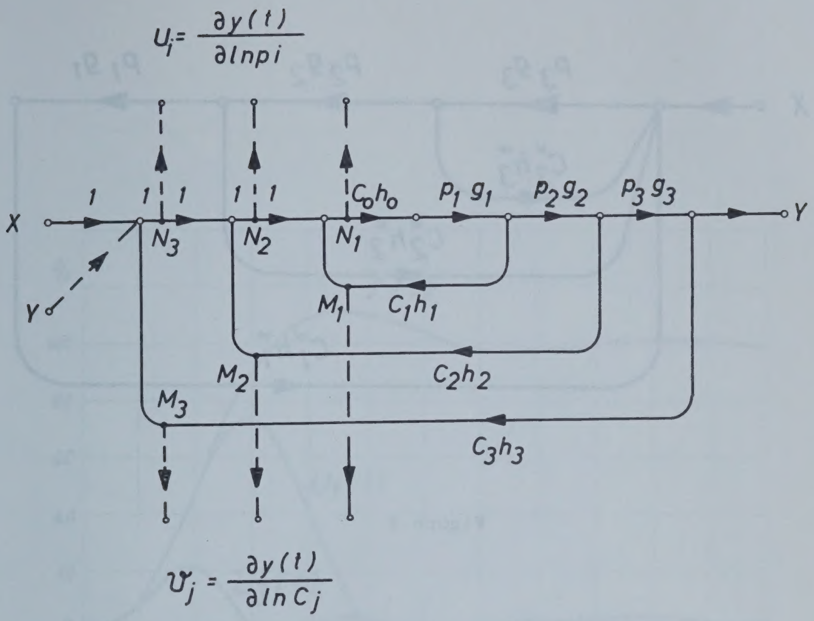

$U_i = \frac{\partial y(t)}{\partial \ln p_i}$
1
1
1
1
1
$C_o h_o$
$P_1 g_1$
$P_2 g_2$
$P_3 g_3$
X
Y
N_3
N_2
N_1
Y
M_1
$C_1 h_1$
M_2
$C_2 h_2$
M_3
$C_3 h_3$
$V_j = \frac{\partial y(t)}{\partial \ln C_j}$

Figure 1

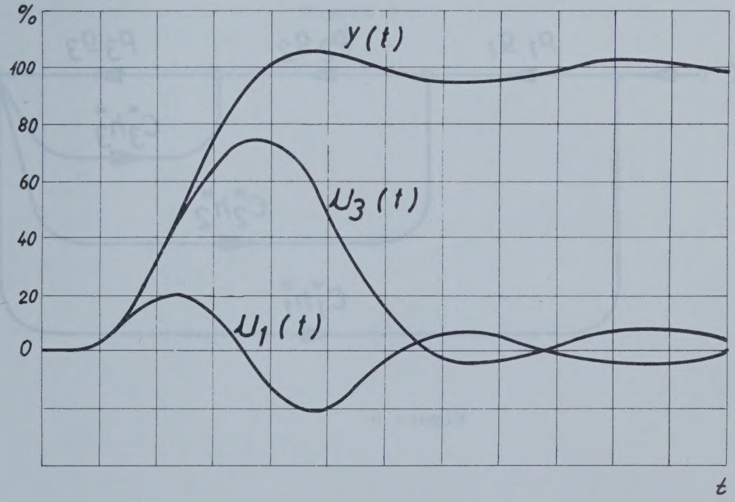
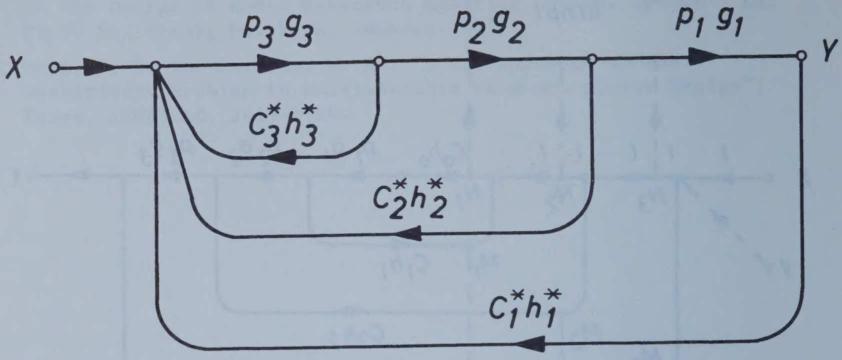
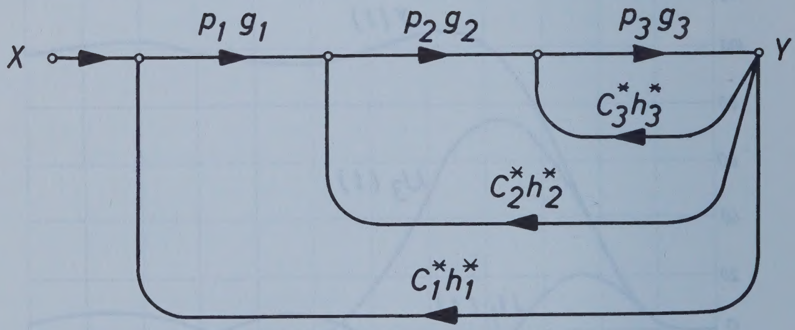


Figure 2



$p_3 g_3$
$p_2 g_2$
$p_1 g_1$
X
Y
$c_3^* h_3^*$
$c_2^* h_2^*$
$c_1^* h_1^*$

Figure 3



$p_1 g_1$
$p_2 g_2$
$p_3 g_3$
X
Y
$c_3^* h_3^*$
$c_2^* h_2^*$
$c_1^* h_1^*$

Figure 4

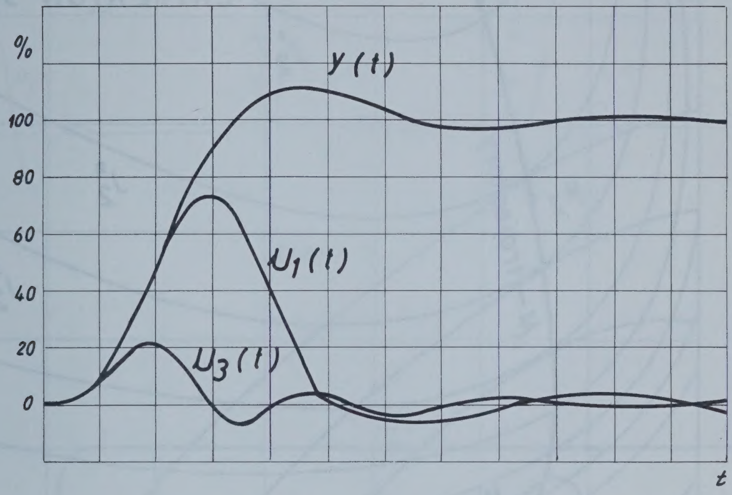


Figure 5

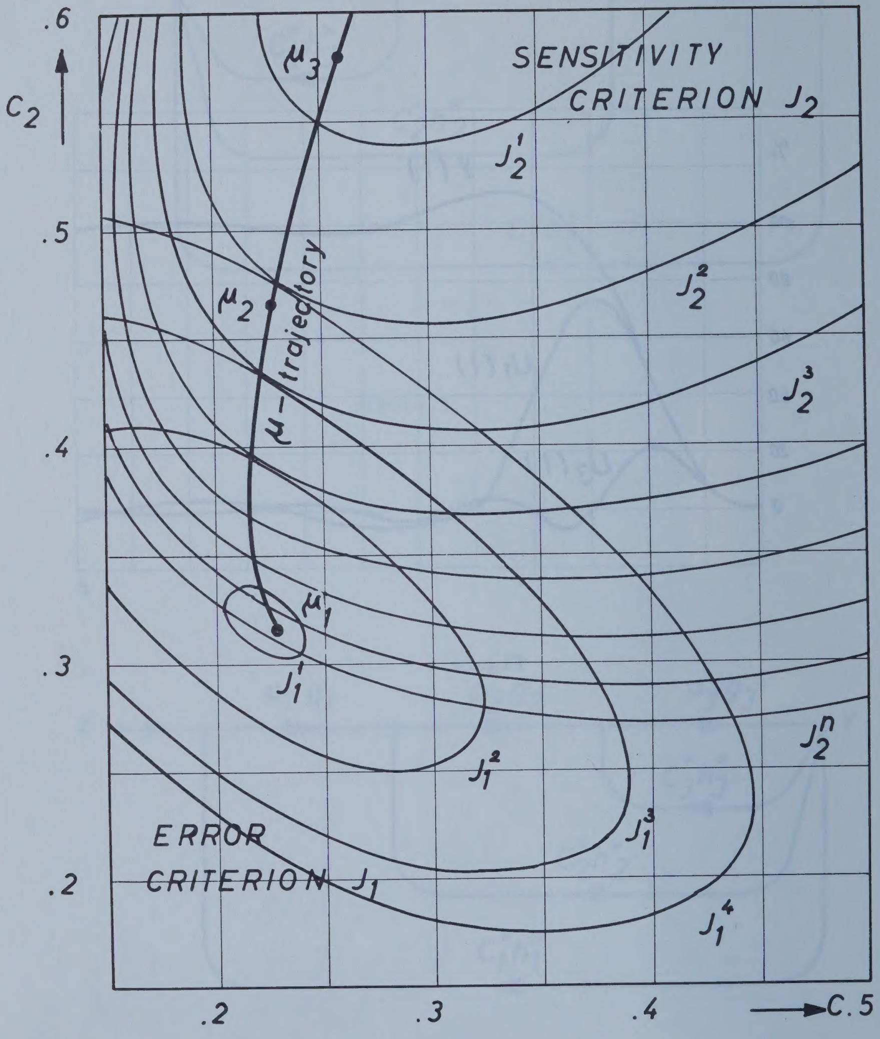

.6
.5
.4
.3
.2
C_2
μ_3
SENSITIVITY
CRITERION J_2
J_2^1
J_2^2
J_2^3
J_2^n
μ_2
μ-trajectory
μ_1
J_1^1
J_1^2
J_1^3
J_1^4
ERROR
CRITERION J_1
.2
.3
.4
C.5

Figure 6

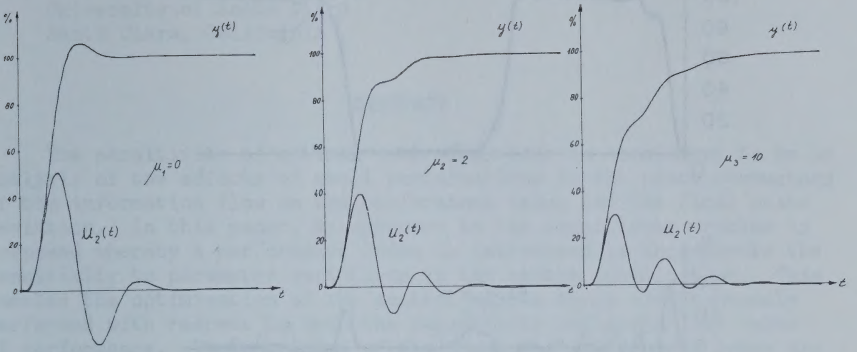


Figure 7

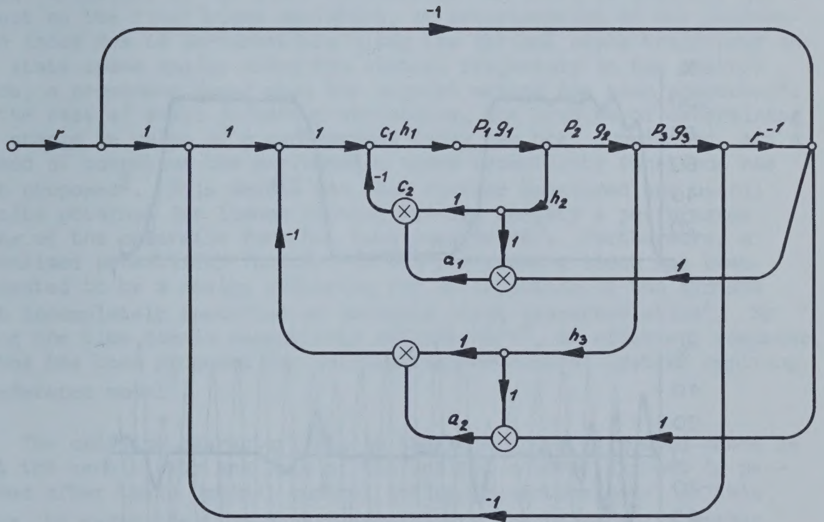

-1
r
1
1
1
c1h1
p1g1
p2g2
p3g3
r^-1
-1
c2
1
h2
1
a1
1
-1
1
h3
1
a2
1
-1

Figure 8

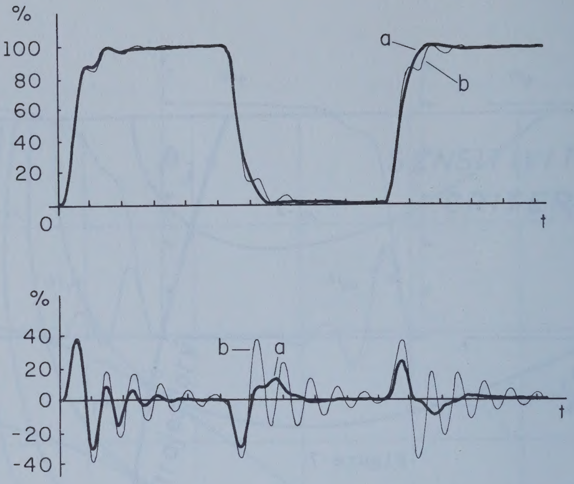


Figure 9

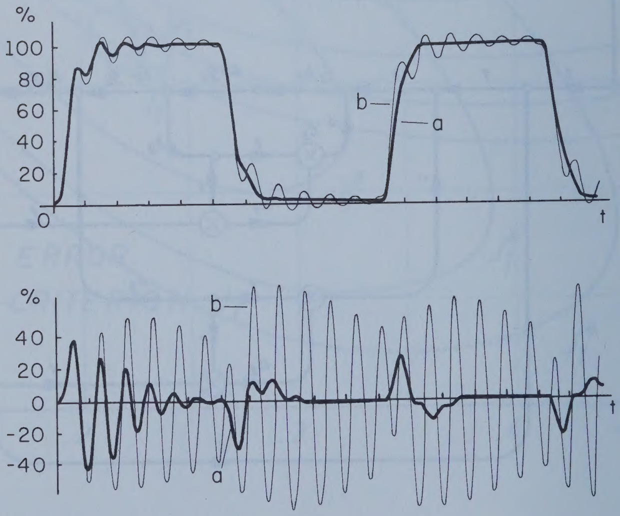


Figure 10

ON THE MINIMIZATION OF SENSITIVITY IN OPTIMAL CONTROL SYSTEMS

D. D. ŠILJAK and R. C. DORF
University of Santa Clara
Santa Clara, California

ABSTRACT

The sensitivity of optimal control systems is considered to be an analysis of the effects of small perturbations in the plant parameters or the information flow on the performance index and the final state deviation. In this paper, an approach to the sensitivity problem is proposed whereby a performance index is introduced to incorporate the sensitivity to parameter variations in the system optimization. This enables the optimization of the control system to be simultaneously performed with respect to both the sensitivity and optimality index of performance. Various forms of the suggested performance index are discussed.

INTRODUCTION

The optimal control action and the corresponding optimal state trajectory of a plant under control are determined. The sensitivity problem may arise due to uncontrolled small perturbations in the information flowing from the plant to controller and vice versa, or small variations of the plant parameters. In order to evaluate the effect on the final state deviation, or deterioration of the performance index due to perturbations along the optimal state trajectory in the state space and/or along the control trajectory in the control space, a procedure based upon the adjoint method has been presented[1]. In the case of small parameter variations, the problem of determining the change in value of a performance index has been considered, and a method of computing the performance index sensitivity functions has been proposed[2]. This method has been further developed and useful results obtained for linear control systems whereby a performance index of the quadratic form has been considered[3]. Furthermore, a normalized sensitivity function of the performance index has been suggested to be a design criterion for optimization of the systems with incompletely specified or variable plant characteristics[4]. By using the time domain sensitivity definition[5,6], an efficient computer method has been proposed for optimization of control systems applying a reference model[7].

The unifying characteristic in the references discussed above is that the sensitivity analysis of the optimal control systems is performed after their optimal control action is synthesized. In this paper, by using the time-domain sensitivity technique[5,8], a rather general index of optimality is introduced wherein both sensitivity and optimality characteristics are incorporated. This enables the synthesis of the optimal control action to be performed simultaneously with respect to sensitivity and optimality requirements of the system optimization. In the following, various forms of the suggested performance index are discussed.

THE OPTIMIZATION PROCEDURE

The dynamic response of a multivariable system may be represented by the matrix differential equation

$$\dot{\underline{x}} = \underline{F}(\underline{x},\underline{u},\underline{p},t) \tag{1}$$

where $\underline{x}$ is the vector of state variables, $\underline{u}$ is the vector of control variables and $\underline{p}$ is the vector of parameters. If the system is represented by the Taylor series expansion about the operating point $\underline{x}_o$, equation 1 may be written as

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{B}\,\underline{u} \tag{2}$$

where $\underline{A}$ is the Jacobian matrix so that $a_{ij} = \frac{\partial f_i}{\partial x_j}$ and $b_{ij} = \frac{\partial f_i}{\partial u_j}$. The solution of the linear vector differential equation is

$$\underline{x}(t) = \underline{\phi}(t)\underline{x}(o) + \int_o^t \underline{\phi}(t-\tau)\,\underline{B}\,\underline{u}(\tau)\,d\tau \tag{3}$$

where $\underline{\phi}(t) = \exp(\underline{A}t)$ and $\underline{x}(o)$ is the initial value of the state vector[9].

The sensitivity of the state variables to variations of the parameters can be represented by the change of the state vector $\wedge\,\underline{x}(t)$ due to a change of a parameter p_k. Alternatively, the variation in the state variables may be written for a small variation of a parameter p_k as $\underline{v}_k = \frac{\partial \underline{x}}{\partial p_k}$, which are called the parameter influence coefficients or the sensitivity coefficients. If there are K parameters that vary and which must be considered, the variation in the state variables must be evaluated for the parameter vector $\underline{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_K \end{bmatrix}$

Then the change of the state vector $\underline{x}$ due to $\underline{p}$ is

$$\frac{\partial \underline{x}}{\partial \underline{p}} = \begin{bmatrix} \frac{\partial x_1}{\partial p_1} & \cdots & \frac{\partial x_1}{\partial p_K} \\ \vdots & & \\ \frac{\partial x_n}{\partial p_1} & \cdots & \frac{\partial x_n}{\partial p_K} \end{bmatrix}$$

$$= \left[\frac{\partial \underline{x}}{\partial p_1}, \ldots, \frac{\partial \underline{x}}{\partial p_K}\right]$$

$$= \left[\ \underline{v}_1, \ldots, \underline{v}_K\ \right]$$

where $\underline{v}_1 = \frac{\partial \underline{x}}{\partial p_1}$. The sensitivity coefficients may be obtained by perturbation techniques, partial derivatives, or time-domain matrices[8].

The sensitivity coefficient $\underline{v}_k(t)$ is a function of time and cannot readily serve as a metric or index since it is not a number but a collection of numbers.

In order to compare one system design with another, some numerical measure of system sensitivity is necessary. In general, an integral function of the sensitivity time response $\underline{v}_k(t)$ is useful and, therefore, a general index is

$$I = \int_0^T k(\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_K)\, dt \; . \tag{4}$$

Some useful and practical specific indices are

$$I = \int_0^T (\underline{v}_1^{\,2} + \lambda_1 \underline{v}_2^{\,2} + \ldots + \lambda_{K-1} \underline{v}_K^{\,2})dt \tag{5}$$

and

$$I = \int_0^T (|\underline{v}_1| + \lambda_1 |\underline{v}_2| + \ldots + \lambda_{K-1} |\underline{v}_K|)dt \tag{6}$$

The optimal control problem is normally taken as the problem of minimizing a given performance index I where

$$I = \int_0^T f(\underline{x}, \underline{u}, t)dt \tag{7}$$

under the constraints on the control vector $\underline{u}$ and the state vector $\underline{x}$. Then the sensitivity of the performance index due to a change in the parameters is evaluated and defined as the sensitivity of the optimal system.

By contrast, it is the purpose of this paper to include the effects of parameter variations in the performance index. Then the minimization of the performance index results in an optimum system simultaneously in the sense of control and sensitivity. A general form of the performance index useful for this purpose is

$$I = \int_0^T f(\underline{x}, \underline{u}, \underline{v}_1, \underline{v}_2, \ldots, \underline{v}_K, t)dt. \tag{8}$$

The optimal closed-loop control law $\underline{u}^o(t)$, which minimizes the integral I, may be obtained by utilizing the modern optimization techniques[10,11]. The optimal control law is then written as

$$\underline{u}^o(t) = \psi(\underline{x}, \underline{v}_1, \ldots, \underline{v}_K, t)$$

or

$$\underline{u}^o(t) = \psi_1(\underline{x}, \underline{p}, t). \tag{9}$$

The optimal control problem including the effect of sensitivity may be formulated in several ways. One standard approach is to utilize a formulation in terms of quadratic functions of the variables. Then the optimal control problem becomes the minimization of the integral

$$I = \int_0^T (\lambda_1\underline{x}^2 + \lambda_2\underline{u}^2 + \lambda_3\underline{v}_1{}^2 + \dots \lambda_{K+2}\underline{v}_K{}^2)dt. \tag{10}$$

Since the sensitivities of several state variables are the most important, an index which weighs $\frac{\partial x_i}{\partial p_k}$ individually is desirable. One such index would be

$$I = \int_0^T (\lambda_1\underline{x}^2 + \lambda_2\underline{u}^2 + \lambda_3(\underline{w}^T\underline{v}_1)^2 + \dots + \lambda_{K+2}(\underline{w}^T\underline{v}_K)^2 dt \tag{11}$$

where $\underline{w}^T$ is the transpose of the weighting vector

$$\underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

In the case where only one parameter is to be considered in the optimal control problem, a new state vector y can be defined so that the sensitivity vector $\underline{v}_1$ can be added or adjoined to the system state variables. Then we have

$$\underline{y} = \begin{bmatrix} \underline{x} \\ \underline{v}_1 \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ \frac{\partial x_1}{\partial p_1} \\ \vdots \\ \frac{\partial x_n}{\partial p_1} \end{bmatrix} \tag{12}$$

Then the optimal control problem may be formulated as the minimization of the integral

$$I = \int_0^T g(\underline{y},\underline{u},t)dt \tag{13}$$

and a quadratic formulation of this problem is

$$I = \int_{0}^{T} (\underline{y}^2 + \lambda\, \underline{u}^2)dt \qquad (14)$$

The optimum control law $\underline{u}^o(t)$, which minimizes the integral I, results in an optimum system which is optimum with respect to both performance and sensitivity. The specific control law obtained for a given problem will depend upon the integral chosen and the values of the weighting factors λ_i. Future research will result in an evaluation of the utility of the various proposed integrals and the optimization approach.

REFERENCES

1. Gavrilović, M., Petrović, R., and Šiljak, D., "Adjoint Method in the Sensitivity Analysis of Optimal Systems", J. Franklin Inst. Vol. 276, No. 1, July 1963: pp 26-38.

2. Dorato, P., "On Sensitivity in Optimal Control Systems", IEEE Trans. on Automatic Control, Vol. AC-8, July 1963: pp 256-257.

3. Pagurek, B., "Sensitivity of the Performance of Optimal Control Systems to Plant Parameter Variations", ibid., Vol. AC-10, April 1965: pp 178-180.

4. Rohrer, R. A., and Sobral, M., Jr., "Sensitivity Considerations in Optimal System Design", Proc. of the Nat'l Elec. Conf., Vol. XX, Paper No. 1830, October 1964; pp 746-747.

5. Meissinger, H. F., "The Use of Parameter Influence Coefficients in Computer Analysis of Dynamic Systems", West. Jt. Comp. Conf., San Francisco, May 1960: pp 181-192.

6. Bihovsky, M. L., "Sensitivity and Dynamic Accuracy of Control Systems", (in Russian) Proc. of Acad. of Sci. USSR, Technical Cybernetics, No. 6, Nov. and Dec., 1964; pp 130-144.

7. Kokotović, P., "Sensitivity Points Method in Investigation and Optimization of Linear Control Systems" (in Russian), ibid., Automation and Remote Control t. 25, No. 12, Dec. 1964; pp 1670-1677.

8. Dorf, R. C., "System Sensitivity in the Time-Domain", Proc. of the Allerton Conf. on Circuit and System Theory, Oct. 1965.

9. Dorf, R. C., "Analysis and Design of Control Systems", Addison-Wesley Pub. Co., Reading, Mass., 1965.

10. Bellman, R., "Adaptive Control Processes: A Guided Tour", Princeton Univ. Press, Princeton, N. J., 1961.

11. Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mischenko, E. F., "The Mathematical Theory of Optimal Processes", John Wiley and Sons, Inc., New York, N. Y., 1962.

MINIMIZATION OF SENSITIVITY WITH STABILITY CONSTRAINTS IN LINEAR CONTROL SYSTEMS*

D. D. ŠILJAK and A. BURZIO
University of Santa Clara
Santa Clara, California

ABSTRACT

This paper presents a method for minimizing the sensitivity to parameter variations in linear control systems under the constraint that the system has a desired degree of relative stability. In the proposed procedure, an integral is established as an index of sensitivity. The integral is then minimized in the parameter plane wherein all the roots of the characteristic equation are placed in evidence. This enables the transient response requirements to be incorporated in the minimization of the sensitivity index, and thus the sensitivity and the system response characteristics can be considered simultaneously in control system design. The method is illustrated by examples.

INTRODUCTION

In the design of linear control systems, both the achievement of a certain degree of stability and the minimum sensitivity of system response to parameter variations are desired. These two requirements, however, are usually in contradiction, and a compromise must be made. This problem has been discussed in the literature which is now reviewed.

By using the frequency domain sensitivity [2-4], it is possible to consider the influence of small parameter changes on the frequency response characteristics, and to minimize the mean squared variation of the system response [5]. Along with the consideration of sensitivity, the standard frequency response technique is used to determine system stability and response characteristics. Furthermore, by using the frequency domain sensitivity, the change of a system parameter can be found which will place a stable system on the verge of instability[6].

If the algebraic methods, which solve for the roots of the system characteristic equation, are applied to the design of a control system, the pole-zero sensitivity [7-9] can be used to perform the sensitivity analysis. A technique has been proposed to compensate a linear control system for both the desired location of system dominant roots and their relative insensitivity to parameter variation [10].

By using the time-domain sensitivity [11], an efficient computer method has been devised to investigate control system sensitivity [12]. The method has been applied to the automatic computer optimization of control systems using a reference model.

---

*This research was partially supported by the National Aeronautics and Space Administration under contract NAS2-2609.

The above-mentioned methods enable the designer to satisfy both stability and sensitivity requirements. The design process, however, consists of interchanging the steps for stability and sensitivity considerations until a satisfactory solution is obtained. Even if such a design process is convergent to a desired set of system performance characteristics, it is a tedious procedure.

This paper, based upon the application of the parameter plane method[1], presents a procedure which enables the stability and sensitivity requirements to be considered simultaneously in control system design. In the procedure, a time domain sensitivity index is minimized within the required stability constraints. Since the parameter plane method places in evidence all the roots of the system characteristic equation, desired transient response characteristics can be incorporated in the minimization of the sensitivity index in the parameter plane. Limitations of the proposed method may arise due to the fact that a limited number of adjustable system parameters can be considered simultaneously.

SENSITIVITY INDEX

In a linear control system subject to an input signal, the error $e = e(t,q_1,q_2,\dots,q_r)$ is a function of the time t and a set of system parameters $q_1,q_2,\dots,q_r$. To investigate how the small changes in a parameter q_i affect the error signal, parameter influence coefficients u_i have been defined[11] as

$$u_i \overset{\Delta}{=} \frac{\partial e}{\partial q_i}, \quad i = 1, 2, \dots, r. \tag{1}$$

For a given set of values of the system parameters $q_1,q_2,\dots q_r$, the coefficient $u_i = u_i(t)$ is a function of time. The shape of the function $u_i(t)$ determines the sensitivity of the error signal to differential changes in the parameter q_i. The function $u_i(t)$ depends also on the input signal which is usually a test signal such as an impulse-, step-, or ramp-function. Therefore, in order for $u_i(t)$ to be used in the sensitivity analysis of a control system, the input signal must be defined. In this paper, a unit impulse is chosen as the input test signal.

The basic weakness of the coefficient u_i as a measure, or index, of system sensitivity is that a function cannot readily serve as a metric of system performance. The purpose of the performance index is to encompass in a single number, or a limited group of such numbers, a quality measure for the performance of the system. Then the system can be optimized on the basis of this number. Therefore, a general index is introduced

$$I = \int_0^\infty F(u_1,u_2,\dots,u_r,t)\, dt \tag{2}$$

which establishes an integral as a measure of the system sensitivity to small parameter variations. A set of parameter values, which minimizes the above integral, provides the least sensitive system according to the defined performance index. In a system design, the choice of the function F should be governed by usefulness in that the particular index which is selected must be a convenient one to use, as well as one which yields practical results.

In this paper, the sensitivity index

$$I_i = \int_o^\infty u_i^2(t)dt \tag{3}$$

is used. The index considers the influence of the small variations in a system parameter q_i. By using Parceval's theorem, the index I_i of equation 3 can be readily expressed as a function of system parameters.

When the test input signal (a unit impulse) is applied to a linear control system, the Laplace transform of the error signal is a rational function of s given as

$$E(s) = \frac{h(s)}{f(s)}, \tag{4}$$

where

$$f(s) = \sum_{k=o}^{m} a_k s^k \tag{5}$$

$$h(s) = \sum_{k=o}^{n} b_k s^k, \quad m \geq n.$$

The coefficients a_k and b_k are functions of the system parameters

$$\begin{aligned} a_k &= a_k(q_1,q_2,\ldots,q_r) \\ b_k &= b_k(q_1,q_2,\ldots,q_r) . \end{aligned} \tag{6}$$

From equation 1 it follows that the Laplace transform $U_i(s)$ of the coefficient $u_i(t)$ is

$$U_i(s) = \frac{\partial E(s)}{\partial q_i} . \tag{7}$$

Using equations 4-7, one obtains the function $U_i(s)$ in the form

$$U_i(s) = \frac{\sum_{k=o}^{m} a_k s^k \sum_{k=o}^{n} \frac{\partial b_k}{\partial q_i} s^k - \sum_{k=o}^{m} \frac{\partial a_k}{\partial q_i} s^k \sum_{k=o}^{n} b_k s^k}{\left[\sum_{k=o}^{m} a_k s^k\right]^2} . \tag{8}$$

Thus, $U_i(s)$ is a rational function in s, and Parceval's theorem can be used to express the integral I_i of equation 3 as

$$I_i = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} U_i(-s)\, U_i(s)\, ds \tag{9}$$

By substituting equation 8 into equation 9 and using the standard integral tables[13], the integral I_i of equation 9 can be calculated in terms of the system parameters to obtain

$$I_i = I_i(q_1, q_2, \ldots, q_r). \tag{10}$$

Thus the minimization of the sensitivity is reduced to a minimization of the function I_i with respect to the system parameters $q_1, q_2, \ldots, q_r$.

It should be noted that the proposed index of sensitivity can be readily applied to any signal in the system other than the error signal. Furthermore, the form of the index can be extended to the case when the sensitivity of several signals with respect to multi-parameter variations is to be considered.

MINIMIZATION PROCEDURE

After the sensitivity index is expressed as a function of system parameters, the minimization procedure can be formally done in the standard manner of equating the partial derivatives of the obtained function to zero and solving the resultant set of equations

$$\frac{\partial I_i}{\partial q_j} = 0, \quad j = 1, 2, \ldots, r, \tag{11}$$

for the values of the parameters. The set of parameters which satisfies simultaneously equations 11 corresponds to a system which is relatively insensitive to small variations in the parameter q_i. Before the solution set of parameters is accepted, the higher-order derivatives should be calculated to verify the minimum of the function I_i.

The formal minimization procedure outlined above suffers from a serious limitation in that the obtained set of parameters may result in an unstable, or even unrealizable, system. Therefore, in the minimization of the sensitivity index, certain stability constraints have to be introduced. The design objective in this paper is to determine values of system parameters which minimize the sensitivity index under the constraint that all the roots of the system characteristic equation lie at certain desired locations. These simultaneous stability and sensitivity considerations are performed in the parameter plane[1].

In the parameter plane method, the characteristic equation of the system is considered in the form

$$f(s) = \sum_{k=o}^{m} a_k s^k = 0. \tag{12}$$

If the complex variable s is substituted by

$$s = -\omega_n \zeta + j\omega_n \sqrt{1-\zeta^2}\ , \tag{13}$$

where ω_n is the undamped natural frequency and ζ is the relative damping coefficient, equation 12 can be solved for two system parameters α and β :

$$\alpha = \alpha(\omega_n, \zeta), \quad \beta = \beta(\omega_n, \zeta) . \tag{14}$$

Parameters α and β are chosen among the parameters $q_1, q_2, \ldots, q_r$ which appear in the coefficients a_k (equation 6).

In the $\alpha\beta$ plane, characteristic curves are plotted using equation 14, and the regions are determined for which all the roots of equation 12 have the damping coefficient ζ greater than a specified value. Furthermore, since the parameter plane diagram places in evidence all the root values, the system response characteristics can be easily observed. Then, if all the parameters $q_1, q_2, \ldots, q_r$, with the exception of α and β, are specified, the performance index I_i is only a function of these two variable parameters. Therefore, the I_i = const. curves can be plotted in the parameter plane along with the diagrams which determine the system stability and response. The values of the parameters α and β, which minimize the sensitivity index inside a desired relative damping region, correspond to a design solution. After the parameter plane diagram is plotted and the I_i = const. curves are introduced, the solution values of α and β are found by inspection. This procedure will be illustrated in the following examples.

APPLICATIONS

Example 1: Consider a control system with the block diagram of Fig. 1 and the transfer functions

$$G_1(s) = 2 \qquad G_2(s) = \frac{K_2}{0.2s^2 + 0.8s + 1}$$

$$G_3(s) = \frac{1}{0.4s + 1} \qquad G_{-1}(s) = \frac{0.5K_{-1}\, s}{0.5s + 1}\ . \tag{15}$$

The design problem consists in choosing values of the parameters K_2 and K_{-1} such that the sensitivity of the system to changes in K_2 is minimum, under the constraint that the damping ratio ζ of all roots of the characteristic equation

$$f(s) = 0.04s^4 + 0.34s^3 + (1.12 + 0.2K_2K_{-1})s^2 + \\ + (1.7 + K_2 + 0.5K_2K_{-1})s + 2K_2 + 1 = 0 \tag{16}$$

are greater than, or at least equal to, 0.45.

If the variable parameters α and β are chosen to be K_2 and K_2K_{-1}, the curve $\zeta = 0.45$ is determined by the equations

$$\alpha = \frac{0.0076\,\omega_n^4 + 0.22\,\omega_n^3 - \omega_n^2 + 2\,\omega_n + 2.4}{0.2\,\omega_n^2 - 0.36\,\omega_n + 1}$$

$$\beta = \frac{-0.008\,\omega_n^5 + 0.06\,\omega_n^4 - 0.15\,\omega_n^3 + 0.22\,\omega_n^2 + 0.18\,\omega_n - 0.5}{0.2\,\omega_n^2 - 0.36\,\omega_n + 1} \tag{17}$$

and is plotted in Fig. 2. The region $\zeta \geq 0.45$ is then determined in the usual fashion[1]. The stability constraint is satisfied by all values of α and β which lie inside the region $\zeta \geq 0.45$.

To find the point of minimum sensitivity to variations in the parameter K_2, let the input test signal be a unit impulse. From equations 4, 8, 15, and 16, one obtains

$$U_i(s) = \frac{0.4s^5 + 0.42s^4 + 1.8s^3 + 3.94s^2 + 4.4s + 2}{[0.04s^4+0.34s^3+(1.12+2K_2K_{-1})s^2+(1.7+K_2+0.5K_2K_{-1})s+2K_2+1]^2} . \quad (18)$$

Equation 18 is used to determine the integral of equation 9 for the system under investigation. Evaluation of the integral of equation 9 for a sufficient number of (α, β) pairs, by means of the standard tables and a digital computer, enables the plotting of the I_i=I=const. curves shown in Fig. 2. It is evident from this diagram that the point M(4.5;0.7), which is the point of tangency of the curves $\zeta = 0.45$ and I = 0.07, corresponds to the parameter values which satisfy the design conditions. These parameters are $K_2 = 0.7$ and $K_{-1} = 6.5$. The complex poles of the closed-loop system, calculated from the curve $\zeta = 0.45$ and equation 13, are $s_{1,2} = -2.5 \pm j5$. The other two poles are real, $s_3 = -0.7$ and $s_4 = -2.7$.

Example 2: Consider again the block diagram of Fig. 1. Let the transfer functions be

$$G_1(s) = 1 \qquad G_2(s) = \frac{2}{0.2s^2 + 0.8s + 1}$$

$$G_3(s) = \frac{0.5(s+1)}{0.2s + 1} \qquad G_{-1}(s) = \frac{K_{-1}\, s}{T_{-1}\, s + 1} . \quad (19)$$

The input test signal is a unit impulse. The design objective is to investigate the sensitivity of the error signal to variations in the parameter T_{-1}.

The characteristic polynomial f(s) and the polynomial h(s) of equations 5 corresponding to the given system are

$$f(s) = 0.04T_{-1}s^4 + (0.04 + 0.36T_{-1})s^3 + (0.36 + 2T_{-1} + 0.4K_{-1})s^2 + \\ + (2 + 2T_{-1} + 2K_{-1})s + 2 \quad (20)$$

$$h(s) = 0.04T_{-1}s^4 + (0.04 + 0.36T_{-1})s^3 + (0.36 + T_{-1} + 0.4K_{-1})s^2 + \\ + (1 + T_{-1} + 2K_{-1})s + 1 .$$

Choosing $\alpha = T_{-1}$ and $\beta = K_{-1}$, and following a procedure identical to that of the previous example, the parameter plane diagram of Fig. 3 can be plotted. The I = const. curves indicate the sensitivity to variations in the parameter T_{-1}.

From Fig. 3, it is not possible to find in the finite $\alpha\beta$ plane a point of minimum sensitivity for a specified damping ration ζ. However, the I = const. curves can be useful in choosing a desired system response which is relatively invariant to changes in the parameter T_{-1}. For example, the two systems corresponding to the two operating points $M_1(0.2;1.56)$ and $M_2(0.072;3.1)$ may both have satisfactory system response characteristics according to the pole-zero configurations of their closed-loop transfer functions (M_1: poles at -0.4, -5, $-4.3 \pm j10.3$, zeros at -1, -5; and M_2: poles at -0.25, -5.4, $-8.6 \pm j20.8$, zeros at -1, -14). The choice of these two systems can therefore be based upon the relative system sensitivity to variations in T_{-1}. The point M_1 is preferable because M_2 corresponds to the maximum sensitivity, as seen in Fig. 3.

CONCLUSION

A method has been presented by which the sensitivity to variations of system parameters can be used as a criterion in the design of linear control systems. The integration of the error sensitivity coefficients over all time provides a satisfactory sensitivity index which can be minimized in the design procedure. Since absolute minimization of the index usually leads to an unstable or even unrealizable system, the index should be minimized subject to certain stability constraints. As illustrated by the two examples, the parameter plane method can be utilized for simultaneous consideration of the sensitivity, stability, and system response characteristics in control system design.

REFERENCES

1. D. D. Šiljak, "Analysis and Synthesis of Feedback Control Systems in the Parameter Plane", Pt. I - Linear Continuous Systems, Pt. II - Sampled-Data Systems, and Pt. III - Nonlinear Systems, IEEE Trans., Pt. II (Application & Industry), vol. 83, No. 75, Nov. 1964; pp 449-473.

2. H. W. Bode, "Network Analysis and Feedback Amplifier Design", D. Van Nostrand Co., Inc., New York, N. Y., 1945.

3. I. M. Horowitz, "Synthesis of Feedback Systems", Academic Press, New York, N. Y., 1963.

4. J. B. Cruz, Jr., and W. R. Perkins, "A New Approach to the Sensitivity Problem in Multivariable Feedback System Design", IEEE Trans. on Automatic Control, AC-9, July 1964; pp 216-223.

5. W. M. Mazer, "Specification of the Linear Feedback System Sensitivity Function", IRE Trans. on Automatic Control, vol. AC-5, No. 2, June 1960; pp 85-93.

6. R. A. Haddad and J. G. Truxal, "Sensitivity and Stability in Multiloop Systems", IEEE Trans. on Automatic Control, vol. AC-9, Oct. 1964; pp 548-551.

7. Hanoch Ur, "Root Locus Properties and Sensitivity Relations in Control Systems", IRE Trans. on Automatic Control, vol. AC-5, No. 1, Jan. 1960; pp 57-65.

8. R. Y. Huang, "The Sensitivity of the Poles of Linear Closed-Loop Systems", AIEE Trans., pt. II (Applications & Industry), Vol. 77, Sept. 1958; pp 182-187.

9. P. Kokotović and D. D. Šiljak, "The Sensitivity Problem in Continuous and Sampled-Data Linear Control Systems by Generalized Mitrovic's Method", IEEE Trans., pt. II (Applications & Industry), vol. 83, No. 74, Sept. 1964; pp 321-324.

10. B. T. Rung and G. J. Thaler, "Feedback Control Systems: Design with Regard to Sensitivity", IEEE Trans., pt. II (Applications & Industry), vol. 83, No. 75, Nov. 1964; pp 360-366.

11. H. F. Meissinger, "The Use of Parameter Influence Coefficients in Computer Analysis of Dynamic Systems", Proc. of the Western Joint Computer Conference, 1960; pp 181-192.

12. P. Kokotović, "Sensitivity Points Method in Investigation and Optimization of Linear Control Systems" (in Russian), Automatika and Telemeckanika, t. 25, No. 12, Dec. 1964; pp 1670-1677.

13. G. C. Newton, L. A. Gould and J. F. Kaiser, "Analytical Design of Linear Feedback Controls", John Wiley & Sons, Inc., New York, N. Y., 1957.

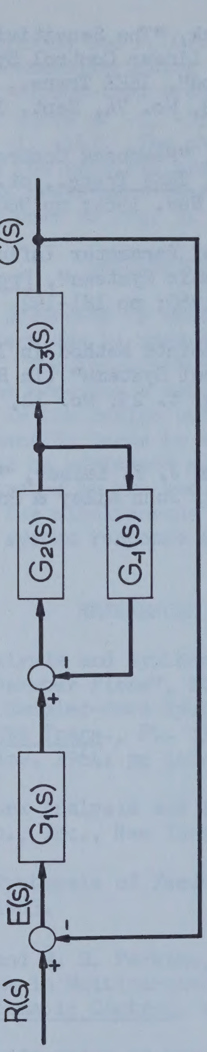

R(s)
E(s)
G₁(s)
G₂(s)
G₄(s)
G₃(s)
C(s)
+
-
+
-

Fig. 1 System Block Diagram

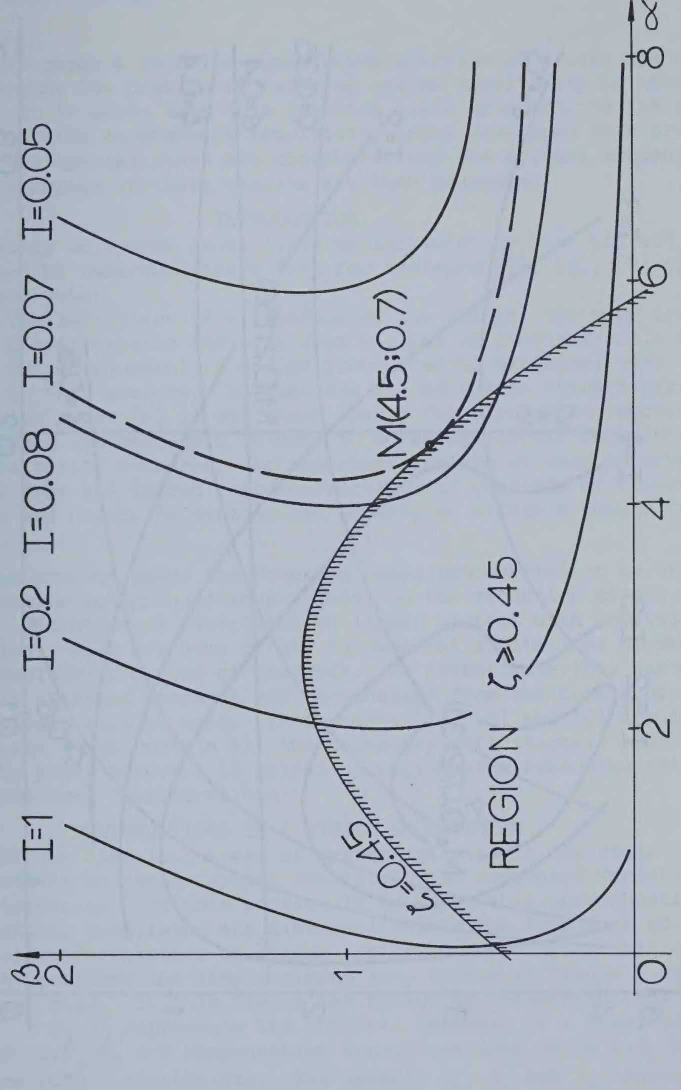


Fig. 2 Parameter Plane Diagram for Example 1.

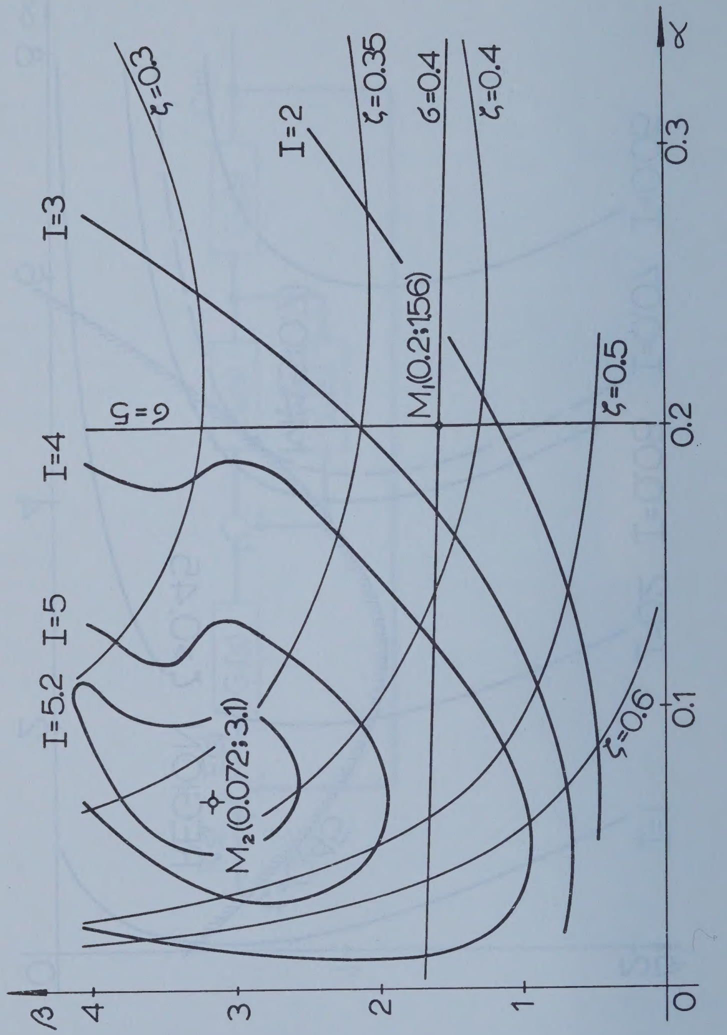


Fig. 3 Parameter Plane Diagram for Example 2.

SOME THEORETICAL LIMITATIONS OF SYSTEM SENSITIVITY REDUCTION[*]

WILLIAM A. PORTER
University of Michigan, Department of Electrical Engineering
Ann Arbor, Michigan

In this paper a function space representation of linear systems is used to examine the problem of reducing system sensitivity by means of feedback. It is shown that this question leads naturally to the problem of minimizing an abstract sensitivity index developed in a previous article. Two special cases are considered and the optimal compensators derived. Examples of these results are then presented.

INTRODUCTION

The study of system sensitivity as initiated by Bode [1] and carried forward by numerous others (see for instance [2], [3], [4] and [5]) is concerned with:

(1) The definition of a 'measure of the change' for some system characteristic (arising from a class of disturbances); and
(2) The development of design procedures to minimize, with respect to this measure, fluctuations in the system characteristic.

In an earlier paper [6] it was shown that a function space representation of linear systems could be used to advantage in the formulation of system sensitivity measures. An important feature of such an approach being that free and forced response sensitivity problems of discrete, continuous and composite systems can be treated within a common framework.

In the present paper the function space representation is used once more to examine design questions related to the reduction of system sensitivity. Attention is restricted to linear systems with emphasis on the question: What are some of the fundamental limitations of the reduction of sensitivity by use of feedback. To investigate this question efficiently standard notation and terminology from the domain of functional analysis must be used. (References [7], [8] and [9] are introductory texts which contain all the necessary definitions.) The development of the paper however, is guided by engineering reasoning rather than mathematical considerations.

PERTURBATIONS IN A SINGLE LOOP SYSTEM

One of the classical areas of system analysis is the study of the use of feedback to reduce system sensitivity to component variations and other disturbances. In this section it is shown that such questions can be meaningfully formulated and discussed within the framework of the function space formulation of linear systems. In studying this class of sensitivity problems the simple closed loop system of Figure 1 will play an important role. In this figure the following transformations are in evidence: $F:B_1 \rightarrow B_2$ represents the internal response of a fixed plant, $G:B_1 \rightarrow B_1$ and $L:B_2 \rightarrow B_1$ are compensation transformations while $K:B_2 \rightarrow B_2$ and $J:B_2 \rightarrow B_2$ are output constraints. The symbols B_1, B_2 and B_3 denote Banach spaces. The system variables are as follows: u and x denotes the system input and useful output respectively, e_4 denotes the observable output

---

[*] The work in this paper was supported by the National Science Foundation under Grant GP-524 and the U.S. Army Research Office-Durham under contract DA-31-124-ARO-D-391.

e_3 denotes the internal (state function space) response, e_2 represents the plant input, while the elements ξ and η denote system disturbances.

Assuming that all transformations are linear, the defining equations for the system are evidently

$$e_3 = \eta + FGe_1$$

$$e_1 = u + \xi - LJe_3.$$

Letting $M = LJ$ this equation set may be reduced to the form

$$(I + MFG)\, e_1 = u + \xi - M\eta$$

$$(I + FGM)\, e_3 = FG(u + \xi) + \eta.$$

Assuming for the moment that $(I + MFG)$ and $(I + FGM)$ are invertible we have

$$e_1 = (I + MFG)^{-1}\, [u + \xi - M\eta] \qquad (1)$$

$$e_3 = (I + FGM)^{-1}\, [FG(u + \xi) + \eta]. \qquad (2)$$

In comparing the sensitivities of open and closed loop systems it is necessary to establish a terminal equivalence between the two system types. In Figure 2 an open loop system which is comparable to the closed loop system of Figure 1 is depicted.[1] Since ξ and η are disturbance elements, the systems of Figures 1 and 2 will be called equivalent if the terminal mapping u→x is the same in both cases. From Equation 2 it follows that the systems are equivalent whenever

$$(I + FGM)^{-1}FG = FQ$$

multiplying through by $(I + FGM)$ and regrouping terms we have

$$FG(I - MFQ) = FQ.$$

Assuming that $(I - MFQ)$ is invertible this equality is satisfied whenever

$$G = Q(I - MFQ)^{-1}. \qquad (3)$$

In Equation 3, F and Q are fixed transformations. This equality then relates G explicitly to the transformation M which remains as the independent compensation transformation of the closed loop system. In addition to Equation 3 several other equalities, summarized by the following lemma will be useful in the following discussion.

Lemma 1. Let Q, G be bounded linear operators on B_1 while $F:B_1 \to B_2$ and $M:B_2 \to B_1$ are bounded linear transformations. Then the condition that the numbers $||FGM||$, $||MFQ||$ and $||FQM||$ are all less than 1 is sufficient (but not necessary) for the identity

(i) $G = Q(I - MFQ)^{-1}$

to imply the identities

(ii) $(I + FGM)^{-1} = (I - FQM)$

(iii) $(I + MFG)^{-1} = (I - MFQ)$

(iv) $(I + FGM)^{-1}\, FG = FQ$

(v) $I - M(I + FGM)^{-1}\, FG = I - MFQ.$

Proof. Identity (v) follows immediately from (iv) which follows from (i). To prove (ii) and (iii) we shall need a standard theorem of the literature (see [10, pg 164]) which states that if $A:B \to B$ and $||A|| < 1$, then the range of I-A is B, $(I - A)^{-1}$ exists, is bounded on B. Furthermore the so called Neumann expansion;

$$(I - A)^{-1} = I + A + A^2 + A^3 + \dots + A^n + \dots \qquad (4)$$

holds, the series converging in norm, and finally

$$||(I - A)^{-1}|| \leq (I - ||A||)^{-1}. \qquad (5)$$

---

[1] The introduction of the transformation Q allows the ideal system to be taken as a compensated form of the plant.

Once it is known that the inverses of the lemma do exist parts (ii) and (iii) may be proved by simply clearing fractions. For example if (iii) holds then

$$\begin{aligned} I &= (I + MFG)(I - MFQ) \\ &= (I - MFQ) + MF[G(I - MFQ)] \\ &= (I - MFQ) + MF[Q] = I. \end{aligned}$$

Remark 1. Where the condition $||MFQ|| < 1$ is satisfied the Neumann expansion provides a synthesis method for the compensation transformation G. Indeed if the transformations Q, M and F are all known then the series

$$G = Q[I + MFQ + MFQMFQ + \ldots]$$

may be synthesized as shown in Figure 3. Furthermore from Equation 4 it is clear that

$$(I - A)^{-1} - [I + A + \ldots + A^n] = A^{n+1}(I - A)^{-1}$$

and hence by Equation 5 it easily follows that

$$||(I - A)^{-1} - [I + A + \ldots + A^n]|| \leq ||A||^{n+1}(1 - ||A||)^{-1}$$

This establishes a means of judging the accuracy of a finitely truncated form of the synthesis ladder in Figure 3.

Let us now return to the investigation of the effects of feedback on system sensitivity. In the system of Figures 1 and 2 the common terminal relation u→x defined by the transformation $KFQ:B_1 \to B_2$ is assumed to define nominal system behavior. We shall consider two types of perturbations from this nominal. First, the disturbance elements $\xi \in B_1$ and $\eta \in B_2$ which have already been included in Equations 1 and 2. Secondly, a bounded linear additive variation δF in the transformation F.

It is clear from Figure 2 that $\bar{e}_2 = Qu$, $\bar{e}_3 = FQu$ and $\bar{x} = KFQu$ are the respective open loop system nominal responses for the input u. The derivations δe_1, δe_3 and δx from nominal are also immediate being given to a first order approximation by:

$$\begin{aligned} \delta e_2 &= Q\xi \\ \delta e_3 &= \delta FQu + FQ\xi + \eta \\ \delta x &= K\delta e_3. \end{aligned}$$

Turning now to the closed loop system of Figure 1 and specifically to Equation 2 we have the perturbation relationship

$$[I + (F + \delta F)GM](\bar{e}_3 + \delta e_3) = (F + \delta F)G(u + \xi) + \eta$$

which reduces to

$$(I + FGM)\delta e_3 = - \delta FGM\,\bar{e}_3 + \delta FGu - \delta FGM\delta e_3 + (F + \delta F)G\xi + \eta.$$

Using the relation $\bar{e}_3 = FQu$ and property (i) of the lemma reduces this expression to

$$\begin{aligned} (I + FGM)\delta e_3 &= \delta F[G - GMFQ]u - \delta FGM\delta e_3 + (F + \delta F)G\xi + \eta \\ &= \delta FQu - \delta FGM\delta e_3 + (F + \delta F)G\xi + \eta. \end{aligned}$$

To a first order effect in the variations $(\delta F, \delta e_3, \xi, \eta)$ the governing expression for δe_3 is therefore

$$\delta e_3 = (I + FGM)^{-1}\,\{\delta FQu + FG\xi + \eta\}. \qquad (7)$$

Consider now the perturbed version of Equation 1,

$$[I + M(F + \delta F)G](\bar{e}_1 + \delta e_1) = u + \xi - M\eta$$

which reduces immediately to

$$[I + MFG]\delta e_1 = -M\delta FG\bar{e}_1 - M\delta FG\delta e_1 + \xi - M\eta.$$

Using Equation 1 (with $\xi, \eta = 0$) to determine $\bar{e}_1$ and properties (iii) and (i) of the lemma, also neglecting second order effects in the deviations, we have

$$[I + MFG]\delta e_1 = -M\delta FQu + \xi - M\eta. \qquad (8)$$

Finally, using properties (ii) and (iii) and recalling that M = LJ, Equations 7 and 8 reduce to the forms

$$\delta e_1 = [I - LJFQ][-LJ\delta FQu + \xi - LJ\eta] \tag{9}$$
$$\delta e_3 = [I - FQLJ][\delta FQu + FQ[I - LJFQ]^{-1}\xi + \eta] \tag{10}$$

which in conjunction with $\delta e_2 = G\delta e_1$ and $\delta x = K\delta e_3$ describe completely the variational behavior of the closed loop system.

Remark 2. Equations 9 and 10 treat the external disturbances, ξ, η on an equal footing with the perturbation δF in the system response function. It is possible to strengthen this tie even more if the disturbance problem is viewed in the following manner. Let the tuplet $(e_1, e_3) \epsilon\ B_1 \times B_2$ be taken as the relevant system response variables and let the tuplet $(u, \xi, \eta) \epsilon\ B_1 \times B_2$ be thought of as the total system input. Define the transformation $(e_1, e_3) = V(u, \xi, \eta)$ by the equations

$$\begin{aligned} e_1 &= I\ u + 0\cdot\xi + 0\cdot\eta \\ e_3 &= T\ u + 0\cdot\xi + 0\cdot\eta. \end{aligned} \qquad (u, \xi, \eta) \epsilon\ B_1 \times B_1 \times B_2 \tag{11}$$

where $T = FQ$. Then, with the meaning of Equations 9 and 10, the deviation δV in V may be written as the operator matrix

$$\delta V = \begin{bmatrix} -(I-MT)M\delta T & (I-MT) & -(I-MT)M \\ (I-TM)\delta T & T & (I-TM) \end{bmatrix} \tag{12}$$

where $\delta T = \delta FQ$. In other words the disturbance (ξ, η) is thought of as being there all the time in a system with nominal response to these signals being multiplication by zero.

As a consequence of these observations it is also possible to measure the total disturbance δV as in the usual parameter variation problem. For example, if B_1, B_2 are the Hilbert spaces H_1, H_2, the sensitivity measures proposed in [6] can be used. Here it is necessary to compute the operator $I + V^*V$ on $H_1 \times H_1 \times H_2$ which, from Equation 11 may be written in the obvious operator matrix form

$$[I+V^*V] = \begin{bmatrix} 2I+T^*T & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

Similarily the normalized perturbation operator is given by

$$\delta V[I+V^*V]^{-1/2} = \begin{bmatrix} -(I-MT)M\delta T[2I+T^*T]^{-1/2} & (I-MT) & -(I-MT)M \\ (I-TM)\delta T[2I+T^*T]^{-1/2} & T & (I-TM) \end{bmatrix} \tag{13}$$

some norm (the usual supremum, the Hilbert-Schmidt or the partitioned-supremum Euclidian introduced in [6]) of which can aid in defining a gross sensitivity function for the system.

A SENSITIVITY MINIMIZATION PROBLEM

The closing remark of the above section indicates how one might define a sensitivity measure for the disturbance problem of Figure 1. This conceptual sketch of the situation has of course passed over some potentially difficult problems in the specification of the various function spaces and the computation of any norm on the operator of Equation 13. The proposed technique does provide a direct relationship between system sensitivity and the compensation transformation M. Thus, while this section considers only simplified cases, it seems altogether feasible that this sensitivity measure can be computed and minimized with respect to a class of 'compensators' with the aid of machine computation.

As a preliminary problem let us focus attention on Equation 10. Setting $\xi, \eta = 0$, $T = FQ$ and $J = I$ this equation becomes

$$\delta e_3 = (I - TM)\delta Tu.$$

An apparent question is: Can M be chosen to make I-TM small while satisfying the constraints $||MT||$, $||TM|| < 1$ of lemma 1? We should also be interested in the form of G, which in the present case is defined by

$$G = (I - MT)^{-1}.$$

To make I-TM small it is evidently both necessary and sufficient to make TM approach the identity operator as closely as possible. To phrase this objective another way, M should be the transformation with largest domain such that I is an extension of TM, that is $TMx = x$, $x \in D(TM)$. Now suppose that $\overline{R(T)}$, the closure of the range of T, is a proper closed subspace of B_2. If I_1, I denote the identities on $\overline{R(T)}$ and B_2 respectively then I extends TM if and only if I_1 extends TM. Hence, without loss of generality, the assumption $\overline{R(T)} = B_2$ can be made.

Theorem 1. Let $T:B_1 \to B_2$ be a bounded one-to-one transformation with dense range. Then T^{-1} exists as a closed densely defined linear transformation. Furthermore:

(1) I extends TM for some linear transformation M if and only if T^{-1} extends M,

(2) There exists no transformation M such that; I extends TM and $||(I-MT)^{-1}|| < \infty$.

Proof. Let M be an operator with domain D(M) in B_2 such that I extends TM. Then if $y \in D(M)$ we have

$$y = T(My)$$

so that y belongs to the range of T and hence $D(T^{-1}) \supset D(M)$. Moreover $T^{-1}y = My$ and hence T^{-1} extends M. Conversely if T^{-1} extends M then $T^{-1}x = Mx$ for $x \in D(M)$ and hence $TMx = TT^{-1}x = x$ holds for $x \in D(M)$ which proves that I extends TM. In other words T^{-1} is the maximal transformation in the class satisfying 1.

Next we shown that if T^{-1} extends M, then I-MT is not invertible. Indeed if $y = x-MTx$ belongs to the range of I-MT, then Tx necessarily belongs to $D(M) \subset D(T^{-1})$ hence

$$MTx = T^{-1}Tx = x$$

Therefore $R(I-MT) = \{0\}$ and I-MT is not invertible.

While part 2 of this theorem rules out the ideal case it is important to note, however, that if $\lambda \neq 0,1$, then $M = \lambda T^{-1}$ satisfies

(a) $TMy = \lambda y$ all $y \in R(T)$

(b) $(I-MT)^{-1} = (1-\lambda)^{-1}I$.

Indeed if $M = \lambda T^{-1}$ then $I-MT = (1-\lambda)I$ has a bounded inverse and if $y = Tx$ belongs to the range of T,

$$TMy = T\, \lambda\, T^{-1}y = \lambda y.$$

Thus assuming that $R(\delta T) \subset R(T)$ and choosing $M = \lambda T^{-1}$ then sensitivity reduction

$$\delta e_3 = (I-TM)\delta Tu = (1-\lambda)\delta Tu$$

may be achieved. Since δT is assumed to be bounded, the ratio $(||\delta e_3||/||u||) \leq |1-\lambda| \cdot ||\delta T||$ can therefore be reduced to any small but finite apriori value by letting $\lambda \to 1$. This reduction in sensitivity is bought at the expense of increasing gain in the forward compensator: $G = (I-MT)^{-1} = (1-\lambda)^{-1}I$.

Assume now that T satisfies the conditions of theorem 1 with the exception that T is not one-to-one. The null space N(T) of T is closed. Suppose that there exists[1] a complementary closed subspace A such that

---

[1] In Hilbert spaces A may be taken as the orthogonal complement of N(T). In Banach spaces, however, this assumption is not to be taken lightly. The reader is referred to the discussion in [10, Section 4.8] of this point.

B_1 can be decomposed into the direct sum $B_1 = N(T) \oplus A$. Then the projection P of B_1 onto A is continuous and T_A, the restriction of T to A, is one-to-one with $R(T_A) = R(T)$. A densely defined pseudo inverse[1], $T^{\dagger}$, for T may be obtained by choosing $T^{\dagger} = T_A^{-1}$. It has the properties: $T^{\dagger}T = P$ and $TT^{\dagger}y = y$, $y \epsilon R(T)$.

If $M:B_2 \to B_1$ is any linear operator, then $x = TMx$, $x \epsilon D(M)$ if and only if $x = T_A PMx$. Hence I extends TM if and only if I extends $T_A PM$ which (by theorem 1) happens if and only if $T^{\dagger}$ extends PM. If $T^{\dagger}$ extends PM, however, then the equality chain

$$MTx = MT_A Px = (I-P)MT_A Px + PMT_A Px = (I-P)MT_A Px + Px$$

holds, which shows that

$$PMT_x = Px.$$

Thus if $y = x - MTx$ is in the range of I-MT then

$$Py = Px - PMTx = 0$$

and hence $R(I-MT) \subset N(T) \neq B_1$. These observations lead to the following corollary.

Corollary. The condition 'that T be one-to-one' of theorem 1 can be deleted provided that a closed direct sum complement exists for N(T). Condition 1 should then read: I extends TM for some linear transformation M if and only if T_A^{-1} extends PM. Condition 2 remains intact.

The case where the compensator M is constrained to be of the form $M = LJ$ leads to analogous results. For instance, with Hilbert spaces and T and J both onto, the choice $L = \lambda T^{\dagger}J^{\dagger}$ produces the results

$$I - TLJ = I - \lambda P$$
$$I - LJT = I - \lambda T^{\dagger}PT$$

where P is the orthogonal projection on the subspace N(J).

Example 1. As an example of our abstract results consider the linear dynamic system of Figure 2 with plant which satisfies the equations

$$\dot{e}_3(t) = A(t)e_3(t) + B(t)e_2(t); \quad e_3(t_o) = 0 \qquad (14)$$

on an interval $\tau = [t_o, t_f]$. Here the functions e_3 and e_2 are n and m tuples of time functions respectively (for example elements of the Hilbert spaces $[L_2(t_o,t_f)]^n$ and $[L_2(t_o,t_f)]^m$ equipped with the usual norm) and the matrices A, B have compatible dimensions. The transformation $e_3 = Fe_2$ may be identified by the equation

$$e_3(t) = \int_{t_o}^{t} \Phi(t,s)B(s)e_2(s)ds \qquad t \epsilon \tau = [t_o, t_f] \qquad (15)$$

where $\Phi(t,s)$ denotes the usual state transition matrix associated with Equation (14).

Without loss of generality assume that $n > m$ and that the columns of B(t) are linearly independent vectors in E^n for (almost) all $t \epsilon [t_o, t_f]$. Then a matrix $B^{\dagger}$ exists such that $B\dagger(t)B(t)$, $t \epsilon [t_o, t_f]$, is the identity on E^m and $B(t)B^{\dagger}(t)$, $t \epsilon [t_o, t_f]$, is the orthonormal projection

---

[1] Obviously $T^{\dagger}$ changes with the choice of the subspace A. In Hilbert spaces by choosing $A = N(T)^{\perp}$ one obtains a $T^{\dagger}$ which carries every $y \epsilon R(T)$ into its unique preimage of minimum norm. In some Banach spaces (for example rotund reflexive space) this latter property can be taken as the criterion for defining $T^{\dagger}$ (see [9, Chapter 4]). Such considerations however lead to a nonlinear pseudo inverse which conflicts with the linearity assumed in Section II.

P_t on the instantaneous range space (column space) of $B(t)$ in E^n. The transformation F defined in equation (15) is one to one and has dense range. Its inverse may be computed by the rule

$$(F^{-1}e_3)(t) = B^{\dagger}(t)[\dot{e}_3(t) - A(t)e_3(t)]; \quad t \epsilon \tau \; e_3 \in R(F). \qquad (16)$$

If $L = \lambda F^{-1}$, $0 < \lambda < 1$, it may be verified directly that $e_2 = Ge_1 = (1-\lambda)^{-1}e_1$ and hence the system equations take the form,

$$e_2(t) = (1-\lambda)^{-1}u(t) - \lambda(1-\lambda)^{-1}B^{\dagger}(t)[\dot{e}_3(t) - A(t)e_3(t)], \quad t \epsilon \tau$$

$$\dot{e}_3(t) = [A(t) + \delta A(t)]e_3(t) + B(t)e_2(t), \quad t \epsilon \tau$$

Eliminating e_2 from this set produces the result

$$[I + \lambda(1-\lambda)^{-1}P_t]\dot{e}_3(t) = [(I + \lambda(1-\lambda)^{-1}P_t)A(t) + \delta A(t)]e_3(t) + (1-\lambda)^{-1}B(t)u(t), \quad t \epsilon \tau.$$

Multiplying through by $(1-\lambda)$ and letting $S_t = I - P_t$ denote the instantaneous projection on the orthogonal complement of the column space of $B(t)$ in E^n this expression becomes

$$[I - \lambda S_t]\dot{e}_3(t) = [I - \lambda S_t]A(t)e_3(t) + (1-\lambda)\delta A(t)e_3(t) + B(t)u(t) \quad t \epsilon \tau$$

Using the fact that $S_t B(t) = 0$, it easily follows that

$$\dot{e}_3(t) = A(t)e_3(t) + B(t)u(t) + (1-\lambda)[I - \lambda S_t]^{-1}\delta A(t)e_3(t), \quad t \epsilon \tau \qquad (17)$$

describes the behavior of the compensated system.

In the case where $m = n$ and $B(t)$ is nonsingular for all $t \in [t_o, t_f]$ (and hence $S_t = 0$) equation 17 clearly shows that the nominal system function has been preserved and that the perturbation has been reduced by the fact $1-\lambda$. In the more general case the identity

$$(1-\lambda)[I - \lambda S_t]^{-1} = S_t + (1-\lambda)P_t$$

when used in Equation 17 shows that the portion of $\delta A(t)$ in the column space of $B(t)$ is reduced by the $(1-\lambda)$ factor while the orthogonal part of this disturbance remains unaffected. Since the feedback signal must pass through the matrix B before reaching the point of disturbance this latter result is intuitive.

A SECOND MINIMIZATION PROBLEM

An important aspect of the sensitivity minimization problem posed in Section II is that the ideal feedback function M may well be (and in fact must be, for the case treated in Section III) an unbounded operator. This was the case in Example 1 as testified to by the presence of the derivative in Equation 14. Thus it is entirely feasible that in keeping δe_3 small the spurious signals δe_1 and δe_2 may be quite large in both a Hilbert space and a point wise sense. From an engineering standpoint this can be a serious defect.

In the system of Figure 1 the variable e_2 represents the actual plant input. Consequently the actual fuel, energy, etc expended by the system is likely to be related to this variable. Continuing the assumption η, $\xi = 0$, it is easily shown that $\delta e_2 = -GM\delta e_3 = G\delta e_1 = -M\delta Tu$, describes the disturbance in e_2. Thus it is natural to attempt to minimize some appropriate combination of the errors

$$\delta e_3 = (I - TM)[\delta Tu]$$

$$\delta e_2 = -M\,[\delta Tu].$$

Since these two errors are in different function spaces, algebraic combinations are not appropriate. Norm combinations, however, such as $||\delta e_3||^2 + ||\delta e_2||^2$ may obviously be considered. In this vein let us restrict attention to Hilbert spaces and consider the functional

$$\epsilon(\delta T) = ||R(I - TM)\delta Tu||^2 + ||KM\delta Tu||^2$$

on δT where R is a bounded invertible operator on H_2 and K is a bounded invertible transformation from H_1 onto H_2. A simple computation shows that $\epsilon(\delta T) = \langle \delta Tu, W(M)\, \delta Tu \rangle$ holds where $W(M)$ is the positive selfadjoint operator on H_2 defined by

$$W(M) = (I - M^*T^*)R^*R(I - TM) + M^*K^*KM.$$

Furthermore, to a first order approximation in δM, it may be shown that

$$W(M + \delta M) - W(M) = \delta M^*\{[T^*R^*RT + K^*K]M - T^*R^*R\} + \{\quad\}^*\delta M.$$

where the coefficient preceding δM is the adjoint of the coefficient succeeding δM^*. Hence if a transformation M_o exists which minimizes $\epsilon(\delta T)$ independent of δTu it must of necessity satisfy

$$[T^*R^*RT + K^*K]M_o = T^*R^*R$$

Since K is by assumption invertible it follows that

$$M_o = (K^*K)^{-1}[I + T^*R^*RT(K^*K)^{-1}]^{-1}T^*R^*R \tag{18}$$

defines the optimal choice of the feedback compensator.

To complete this development, the following equalities will be helpful:

$$\text{(i)}\quad I-(K^*K)^{-1}[I+T^*R^*RT(K^*K)^{-1}]^{-1}T^*R^*RT = [I+(K^*K)^{-1}T^*R^*RT]^{-1}$$

$$\text{(ii)}\quad I-T(K^*K)^{-1}[I+T^*R^*RT(K^*K)^{-1}]^{-1}T^*R^*R = (I+T(K^*K)^{-1}T^*R^*R]^{-1}$$

$$\text{(iii)}\quad [I+R^*RT(K^*K)^{-1}T^*]^{-1}R^*R = R^*R[I+T(K^*K)^{-1}T^*R^*R]^{-1}$$

$$\text{(iv)}\quad [I+(K^*K)^{-1}T^*R^*RT]^{-1}(K^*K)^{-1}T^*R^*R = (K^*K)^{-1}T^*R^*R[I+T(K^*K)^{-1}T^*R^*R]^{-1}$$

The inverses of the form $(I + A)^{-1}$ in these equalities exists because the operator A is self-adjoint and positive definite. Each of these equalities may be verified directly by clearing fractions (as suggested in the proof of lemma 1). Although $||A|| > 1$ a modified Neumann expansion can also be defined for $(I + A^{-1})$ (see [11]).

Now that the form of M_o is known the companion compensator G_o is defined by the formula

$$G_o^{-1} = I - M_oT = I - (K^*K)^{-1}[I + T^*R^*RT(K^*K)^{-1}]^{-1}T^*R^*RT$$

Thus as a consequence of equality (i) it follows that

$$G_o = I + (K^*K)^{-1}T^*R^*RT \tag{19}$$

The form of $W(M_o)$ is also of interest. Equality (ii) is an intermediate form of the identity

$$I - TM_o = [I + T(K^*K)^{-1}T^*R^*R]^{-1}.$$

Using the fact that $(I - TM_o)^* = [I + R^*RT(K^*K)^{-1}T^*]^{-1}$ and identify (iii) it follows that

$$(I - TM_o)^*R^*R(I - TM_o) = R^*R[I + T(K^*K)^{-1}T^*R^*R]^{-2}. \tag{20}$$

Using equality (iv) it can likewise be shown that

$$M_o^*K^*K\, M_o = R^*RT(K^*K)^{-1}T^*R^*R[I + T(K^*K)^{-1}T^*R^*R]^{-2}. \tag{21}$$

Thus adding Equation 20 and 21 the operator $W(M_o)$ may be identified as

$$W(M_o) = R^*R[I + T(K^*K)^{-1}T^*R^*R]^{-1} \tag{22}$$

Remark 3. It is helpful to examine a special case of these results. In particular the case where $R = kI$ and K is of the form $K = (I - R)V$ where $0 < k < 1$ and V is an unitary operator from H_1 onto H_2. Then $K^*K = (1-k)^2\, I$ and $R^*R = k^2I$. Consequently the optimal form of M_o becomes

$$M_o = k^2/(1-k)^2\{I + [k^2/(1-k)^2]T^*T\}^{-1}T^*, \tag{23}$$

the companion compensator G_o takes the form

$$G_o = I + [k^2/(1-k)^2]T^*T \tag{24}$$

and the operator $W(M_o)$ becomes

$$W(M_o) = k^2\{I + [k^2/(1-k)^2]TT^*\}^{-1}.$$

If the transformation M_o is written in the equivalent form

$$M_o = T^*\{[(1-k)^2/k^2]\, I + TT^*\}^{-1}$$

the limiting case as k→1, namely $T^*(TT^*)^{-1} = T^\dagger$, can be recognized. The limit k→0, namely $M_o = 0$, is also apparent from the equivalent limit $\epsilon(\delta T) \to ||M\delta Tu||^2$.

Example 2. In this example the solution of the latter sensitivity problem is shown to have implications which are not immediately evident from the abstract form. To do this in a fairly simple setting the system of Example 1 will be considered. With the assumptions: $B(t) = I$, $t \in [t_o,t_f]$ and $H_1 = H_2 = [L_2(t_o,t_f)]^n$ transformation $e_3 = Te_2$ is defined by Equation 14. The operators R and K will be taken as kI and (1-k)I respectively. Thus, Equations 23 and 24 designate the compensators of interest.

To synthesize M_o or G_o one must first determine T*. In the present case it is easily shown that T* is defined by

$$(T^*x)(t) = \int_t^{t_f} \Phi^*(t_f,s)x(s)ds \qquad t \in [t_o,t_f] \tag{25}$$

which may be also written as

$$(T^*x)(t) = \Psi(t,t_o) \int_{t_o}^{t_f} \Psi(t_o,s)x(s)ds - \Psi(t,t_o) \int_{t_o}^{t} \Psi(t_o,s)x(s)ds \qquad t \in [t_o,t_f] \tag{26}$$

where $\Psi(t,s) = \Phi^*(s,t)$. From Equation 25 it is clear that T* is a pure predictor and hence in the general context of this example completely nonrealizable. Equations 23 and 24 show that this nonrealizability is passed along to the compensators G_o and M_o. Thus although the construction of G_o, M_o and $W(M_o)$ is straightforward its principal value is as a benchmark for comparison with realizable compensators.

Remark 4. Equation 26 shows that if the quantity

$$\xi_o = \int_{t_o}^{t_f} \Psi(t_o,s)x(s)ds$$

could be computed than T* could be realized by the system

$$\dot{z}(t) = -A^*(t)z(t) - x(t); \; z(t_o) = \xi_o, \quad t \in [t_o,t_f]$$

where Ψ is the transition matrix of this latter system and $z = T^*x$. Note also that since $e_3 = \overline{e}_3 + \delta e_3$ and $\overline{e}_3 = Tu$, it is only the variation δe_3 that must be predicted. In specific cases where $\overline{u}$ is fixed and sufficient a priori information about this disturbance exists it follows that the predictive nature of T* and (and hence M_o, G_o) can be (perhaps approximately) realized.

A classic case in point is that of the ballistic missile guidance system. In this case the elements x, u denote deviations from nominal trajectory and nominal thrust profile respectively. The times t_o, t_f represent the beginning of free fall and target impact. The system disturbance arises from spurious atmospheric effects prior to t_o and takes the form $\delta e_3(t) = \Phi(t,t_o)\delta x^o$, $t \in [t_o t_f]$ during the true period of interest. For signals of this form equation 25 can obviously be realized by sampling x at time t_o and precomputing the matrix

$$\int_t^{t_f} \Phi^*(t_f, s)\Phi(s,t_o)ds.$$

Remark 5. The two examples point out an interesting hueristic aspect of the problem. Notice that the transformation T of Equation 14 represents a system with 'inertia'. Hence in Example 1 it is necessary to use an instantaneous unbounded feedback compensator to overcome the 'sluggishness' of the system. Such a compensator turns out to be on the 'boundary of realizability'. Indeed, in Example 2 it is seen that by including a small (i.e., $1-k \approx 0$) measure of the action of the compensation in the optimization criteria the compensator is forced into the predictor category.

CONCLUSIONS

A method has been suggested for specifying compensators which maintain system terminal invariance while reducing system sensitivity. Optimal compensators are derived for two special cases. The examples presented were taken from the domain of continuous time differential systems. However, because of the abstract form of the results the methods apply to discrete, composite and distributive systems as well.

REFERENCES

1. Bode, H.W., "Network Analysis and Feedback Amplifier Design", New York, Van Nostrand, 1945.
2. Horowitz, I., "Synthesis of Feedback Systems", Academic Press, New York, 1963.
3. Cruz, Jr., J.B. and Perkins, W.R., "A New Approach to the Sensitivity Problem in Multivariable Feedback System Design", IEEE Trans. on Automatic Control, Vol. AC-9, July 1964.
4. Rohrer, R.A. and Sobral, Jr., M., "Sensitivity Considerations in Optimal System Design", IEEE Trans. on Automatic Control, Vol. AC-10, January 1965.
5. Tomavic, R., "Sensitivity Analysis of Dynamic Systems", McGraw-Hill Book Company, Inc., New York, 1963.
6. Porter, W.A., "Sensitivity Problems in Linear Systems", IEEE Trans. on Automatic Control, Vol. AC-10, July 1965.
7. Simmons, G.F., "Topology and Modern Analysis", McGraw-Hill Book Company, Inc., New York, 1963.
8. Vulik, B.Z., "Introduction to Functional Analysis for Scientists and Technologists", Addison-Wesley Publishing Compnay Inc., 1963.
9. Porter, W.A., "Modern Foundations of System Engineering", New York, Macmillian Company, 1965.
10. Taylor, A.E., "Introduction to Functional Analysis", John Wiley and Sons, Inc., New York, 1958.
11. Altman, M., "Approximation Methods in Functional Analysis", Lecture Notes, California Institute of Technology, 1959.

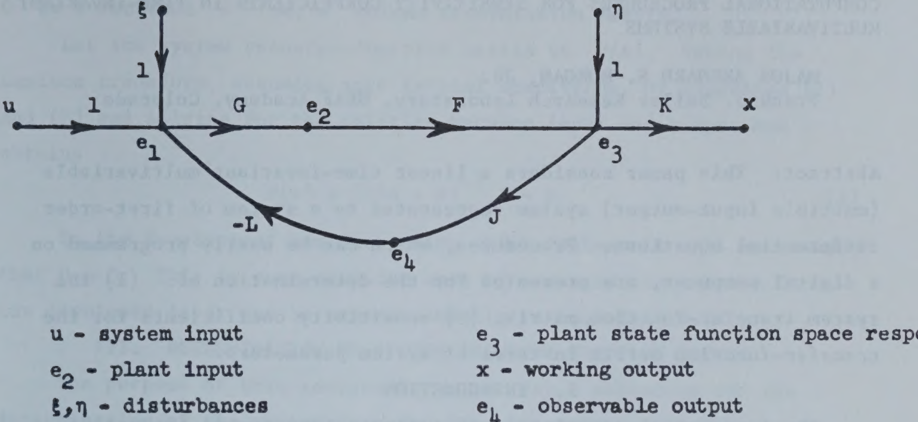

ξ
η
1
1
u
1
G
e2
F
K
x
e1
e3
-L
J
e4
u - system input
e2 - plant input
ξ,η - disturbances
e3 - plant state function space resp
x - working output
e4 - observable output

Figure 1. A Simple Feedback System

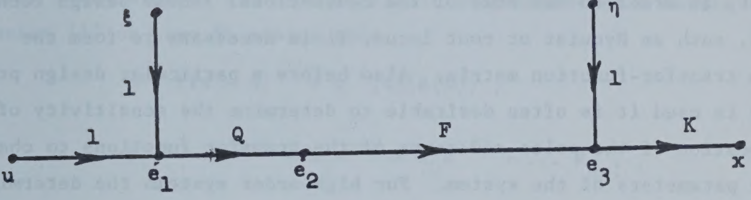

ξ
η
1
1
1
Q
F
K
u
e1
e2
e3
x

Figure 2. A Simple Open Loop System

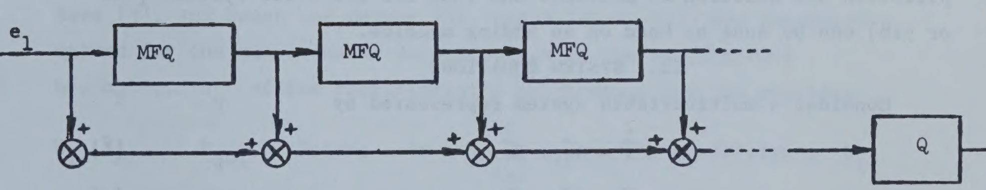

e1
MFQ
MFQ
MFQ
+
+
+
+
+
+
+
Q

Figure 3. The Synthesis of $Q(I-MFQ)^{-1}$

COMPUTATIONAL PROCEDURES FOR SENSITIVITY COEFFICIENTS IN TIME-INVARIANT MULTIVARIABLE SYSTEMS

MAJOR BERNARD S. MORGAN, JR.
Frank J. Seiler Research Laboratory, USAF Academy, Colorado

Abstract: This paper considers a linear time-invariant multivariable (multiple input-output) system represented by a system of first-order differential equations. Procedures, which can be easily programmed on a digital computer, are presented for the determination of: (1) the system transfer-function matrix; (2) sensitivity coefficients for the transfer-function matrix in terms of system parameters.

I. INTRODUCTION

The fundamental description of a physical system is the set of equations based on physical laws that describe its behavior. For linear time-invariant multivariable (multiple input-output) systems this description frequently is a system of differential equations. However, in order to use some of the conventional linear design techniques, such as Nyquist or root locus, it is necessary to form the system transfer-function matrix. Also before a particular design procedure is used it is often desirable to determine the sensitivity of the location of the poles and zeros of the transfer functions to changes in the parameters of the system. For high order systems the determination of the transfer-function matrix from the system of differential equations is a tedious algebraic process and almost impossible when the transfer functions are to be written in terms of the system parameters. It is therefore highly desirable to obtain procedures suitable for programming on a digital computer to accomplish these tasks. Such procedures are presented in this paper. They rely only on the multiplication and addition of matrices and thus for low order systems (4th or 5th) can be done by hand on an adding machine.

II. SYSTEM EQUATIONS

Consider a multivariable system represented by

$$\dot{\vec{x}} = A\vec{x} + B\vec{u} , \tag{1}$$

$$\vec{y} = C\vec{x} + D\vec{u} , \tag{2}$$

where $\vec{u}$ is an m dimensional input vector; $\vec{x}$ is a q dimensional state vector; $\vec{y}$ is an n dimensional output vector; A is a constant q order differential transition matrix; B is a constant q row, m column input matrix; C is a constant n row, q column output matrix;

D is a constant n row, m column transmission matrix.

Let the system transfer-function matrix be P(s). Taking the Laplace transform, assuming zero initial conditions, of equations (1) and (2) and solving for the relation between input and output one obtains

$$P(s) = C[Is - A]^{-1} B + D \tag{3}$$

In the development and discussion which follows it will be assumed that $D \equiv 0$. This is done only for ease of presentation and the formulas developed later can be extended easily to include $D \neq 0$.

III. DETERMINATION OF SYSTEM TRANSFER-FUNCTION MATRIX

The purpose of this section is to present a procedure for the determination of the system transfer-function matrix, represented by equation 3. This procedure forms the basis for later discussions on parameter variations and has the advantage of being easily programmed on a digital computer.

In a discussion of the characteristic polynominal of a matrix Gantmacher [1] on page 84 states that

$$[Is - A]^{-1} = g^{-1}(s)[R(s)] , \tag{4}$$

where

$$R(s) = Is^{q-1} + R_1 s^{q-2} + R_2 s^{q-3} + \ldots R_{q-1} , \tag{5}$$

and

$$g(s) = s^q - h_1 s^{q-1} - h_2 s^{q-2} - \ldots - h_q . \tag{6}$$

R(s) is called the adjoint matrix [1]. Equation (6) defines the characteristic polynominal of the system, i.e.,

$$g(s) = \det [Is - A] . \tag{7}$$

D.K. Faddeev on page 260 [2] (see also Gantmacher [1], page 87, Bass [3], and Zadeh and Desoer [4], pages 302 to 310) has suggested a method for the simultaneous determination of the coefficients $h_1, h_2 \ldots h_q$ of the characteristic polynominal and the matrices $R_1, R_2, \ldots R_{q-1}$. Define $\operatorname{tr} A = \sum_{i=1}^{q} a_{ii}$. The coefficients h_1 and the matrices R_j can be computed from the following formulas:

$$A_1 = A \qquad h_1 = \mathrm{tr}\, A_1 \qquad R_1 = A_1 - h_1 I \tag{8}$$

$$A_2 = AR_1 \qquad h_2 = \frac{1}{2}\,\mathrm{tr}\, A_2 \qquad R_2 = A_2 - h_2 I \tag{9}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$A_{q-1} = AR_{q-2} \qquad h_{q-1} = \frac{1}{q-1}\,\mathrm{tr}\, A_{q-1} \qquad R_{q-1} = A_{q-1} - h_{q-1} I \tag{10}$$

$$A_q = AR_{q-1} \qquad h_q = \frac{1}{q}\,\mathrm{tr}\, A_q \qquad R_q = A_q - h_q I = 0 \tag{11}$$

Equation (11) indicates that R_q should equal zero. This result provides a check on the computation of $R_1 \dots R_{q-1}$, and $h_1 \dots h_q$. It is of interest to note that equation (11) provides a method for computing the inverse of A if the inverse exists, i.e., $h_q \neq 0$. For replacing A_q by AR_{q-1} in the expression for R_q in equation (11) and solving one obtains

$$A^{-1} = (h_q)^{-1} R_{q-1} \; . \tag{12}$$

Now placing equations (4), (5) and (6) in (3) one obtains

$$P(s) = g^{-1}(s) \sum_{i=0}^{q-1} M_i \, s^{q-1-i} \quad , \tag{13}$$

where

$$M_i = CR_i B \; , \tag{14}$$

and

$$R_o = I \; . \tag{15}$$

Therefore once the matrices $R_1 \dots R_{q-1}$ and coefficients $h_1 \dots h_q$ have been found, the matrices M_i and $g(s)$ can be easily computed and the transfer-function matrix, $P(s)$, can be determined.

IV. SENSITIVITY OF P(s) WITH RESPECT TO SYSTEM PARAMETERS

Sensitivity of Characteristic Roots

When the elements of the matrix A change the coefficients of the characteristic polynomial and the roots of the characteristic equation, i.e., the pole locations of the transfer-function matrix, change. This subsection presents formulas for the differential change in any of the elements of A. Once this change is known sensitivity functions can be easily determined. It is assumed that all the roots $s_1, s_2, \dots s_q$ of $g(s) = 0$ are distinct.

An expression for the change, ds_i, in any root s_i for a change in the parameters of A is

$$ds_i = (\mathrm{tr}\, R(s_i))^{-1} \, [R(s_i)]*[dA] \tag{16}$$

where $R(s_i)$ is the adjoint matrix, equation (5) evaluated at s_i; dA is the differential change in the matrix A. The asterisk indicates the inner product of two matrices, that is

$$A*B = \underline{a}_{1\cdot}\vec{b}_1 + \underline{a}_{2\cdot}\vec{b}_2 + \dots \tag{17}$$

where $\underline{a}_{i\cdot}$ is the i^{th} row of A and $\vec{b}_i$ is the i^{th} column of B.

Equation (16) is obtained by starting with a formula of Jacobi [5] for the change ds_i in terms of dA. (Jacobi's formula can also be found in Bellman [6] on page 63 although Bellman's derivation requires that A be symmetric. Jacobi's formula is also derived by Faddeev [2] on page 229.) Jacobi's formula is

$$ds_i = \underline{t}_{i\cdot} \, [dA] \, \vec{p}_i \, , \tag{18}$$

where the $\vec{p}_i$ are characteristic vectors and the $\underline{t}_{i\cdot}$ are characteristic rows. The matrix P of characteristic vectors and the matrix T of characteristic rows are related by

$$P^{-1} = T \, , \tag{19}$$

such that the following is true,

$$s_i\vec{p}_i = A\,\vec{p}_i \, , \tag{20}$$

$$s_i\,\underline{t}_{i\cdot} = \underline{t}_{i\cdot}A \, . \tag{21}$$

Equation (18) can also be written as

$$ds_i = \vec{p}_i\,\underline{t}_{i\cdot} * dA \, . \tag{22}$$

In order to obtain equation (16) one must show that

$$\vec{p}_i\,\underline{t}_{i\cdot} = (\mathrm{tr}\, R(s_i))^{-1} \, [R(s_i)] \, . \tag{23}$$

This is done by first showing that

$$\lim_{s\to s_i} (s-s_i)[Is-A]^{-1} = \lim_{s\to s_i} (s-s_i) \left[Is - P \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_q \end{bmatrix} T\right]^{-1} = \vec{p}_i\,\underline{t}_{i\cdot} \tag{24}$$

and

$$\lim_{s\to s_i} (s-s_i)\, g^{-1}(s)[R(s)] = (g'(s_i))^{-1}[R(s_i)] \, . \tag{25}$$

Then one shows that

$$g'(s_i) = \mathrm{tr}\,[R(s_i)] \, . \tag{26}$$

However

$$\lim_{s\to s_i} (s-s_i)]Is-A]^{-1} = \lim_{s\to s_i} (s-s_i)g^{-1}(s)[R(s)] \quad . \tag{27}$$

Thus equation (23) is obtained by using equations (24), (25) and (26) in equation (27). The formula expressed by equation (26) can be obtained by using equations (8), (9), and (10) in equation (5). That is

$$\mathrm{tr}\ [R(s)] = \mathrm{tr}\ [Is^{q-1} + R_1 s^{q-2} + \ldots R_{q-1}] \ , \tag{28}$$

$$\mathrm{tr}\ [I] = q \quad , \tag{29}$$

$$\mathrm{tr}\ [R_1] = \mathrm{tr}\ [A_1 - h_1 I] = h_1 - qh_1 = -\ (q-1)\ h_1 \ , \tag{30}$$

$$\vdots$$

$$\mathrm{tr}\ [R_{q-1}] = \mathrm{tr}\ [A_{q-1} - h_{q-1} I] = (q-1)h_{q-1} - qh_{q-1} = -h_{q-1} . \tag{31}$$

Thus

$$\mathrm{tr}\ [R(s)] = qs^{q-1} - (q-1)\ h_1\ s^{q-2} - \ldots - h_{q-1} \ , \tag{32}$$

which is g'(s).

Finally it is of interest to comment on equation (23). Since tr $R(s_i)$ is a nonzero scalar it means that the columns of $R(s_i)$ are proportional to the characteristic vector $\vec{p}_i$ and the rows are proportional to the characteristic row $\underset{\smile}{t}_i$.

This observation and the relationship $\underset{\smile}{t}_i\ \vec{p}_i = 1$ permits one to obtain a particular characteristic vector and the corresponding characteristic row from $R(s_i)$. One takes as the characteristic vector the first nonzero column in $R(s_i)$ and as the characteristic row the corresponding first nonzero row appropriately normalized so that $\underset{\smile}{t}_i\ \vec{p}_i = 1$.

Sensitivity of Coefficients of Characteristic Polynomial

This subsection presents formulas for the differential change dh_p in the coefficients of the characteristic polynomial, g(s), for a differential change in the parameters of the matrix A. These formulas are

$$dh_1 = I{*}dA \tag{33}$$

$$dh_2 = R_1{*}dA \tag{34}$$

$$\vdots$$

$$dh_p = R_{p-1}{*}dA \tag{35}$$

These formulas are obtained by first determining an expression for

ds_i in terms of the changes in the coefficients of the characteristic polynomial. This can be done by applying Newton's formula for the approximation to the roots of a polynomial to the characteristic polynomial $g(s)$. If this is done one obtains

$$ds_i = (g'(s_i))^{-1}(dh_1 s_i^{q-1} + dh_2 s_i^{q-2} + \dots dh_q) \tag{36}$$

where $g'(s_i)$ is the derivative of $g(s)$ with respect to s, evaluated at s_i, and a dh_i is the differential change in an h_i. Then using equations (16), (26) and (36) one writes

$$ds_i g'(s_i) = [R(s_i)]*dA = dh_1 s_i^{q-1} + \dots + dh_q \; . \tag{37}$$

After some manipulation and using equation (37) for $i = 1 \dots q$ one obtains

$$[V] \begin{bmatrix} I \\ R_1 \\ \cdot \\ \cdot \\ \cdot \\ R_{q-1} \end{bmatrix} *[dA] = [V] \begin{bmatrix} dh_1 \\ dh_2 \\ \cdot \\ \cdot \\ dh_q \end{bmatrix} , \tag{38}$$

where det $[V]$ is the Vandermonde determinant. Since the matrix $[V]$ is nonsingular for distinct roots, the formulas dh_p follow from (38) by multiplying both sides by the inverse of $[V]$.

Sensitivity of Coefficients of Numerator Polynomials of Transfer-Function Matrix

This subsection presents formulas for the effects of changes in the matrices A, B, and C on the coefficients of the numerator polynomials of any of the transfer functions of the transfer-function matrix. Once these changes are known then the changes in the location of the roots of the numerator polynomials, zeros of any transfer function, can be determined using equation (36).

From equation (14) the coefficient s^{q-1-i} is M_i. Thus a differential change in M_i can be written as

$$dM_i = [dC]R_i B + C[dR_i]B + CR_i[dB] \; , \qquad i = (0, 1, \dots q-1) \; . \tag{39}$$

From equations (8), (9) and (10)

$$dR_i = dA_i - dh_i I \qquad i = (1, 2, \dots q-1) \tag{40}$$

where for $i=0$, $R_0 = I$ so $dR_0 = 0$. Expressions for dh_i are given in equations (34) through (36) and

$$dA_i = [dA]R_{i-1} + A[dR_{i-1}] \ . \tag{41}$$

As an example of the determination of dR_i take $i = 2$ so

$$dR_2 = dA_2 - dh_2 \ I \ . \tag{42}$$

From equations (41) and (9)

$$dA_2 = [dA]R_1 + A[dR_1] \ . \tag{43}$$

However

$$dR_1 = dA_1 - dh_1 I = dA - dh_1 I \ . \tag{44}$$

So using equations (34) and (35) for dh_1 and dh_2 respectively and combining (42), (43) and (44) one obtains

$$dR_2 = [dA]R_1 + A[dA - (I*dA)I] - (R_1*dA) \ I \ . \tag{45}$$

Successive applications of equations (40) and (41) permit the determination of each dR_i and thus each dM_i.

Acknowledgement: The author wishes to thank Professor Elmer G. Gilbert of the University of Michigan for his comments on the work reported in this paper.

V. REFERENCES

[1] F.R. Gantmacher, The Theory of Matrices, Vol I, Chelsea Publishing Company, New York, N.Y., 1959.

[2] D.K. Faddeev and V.N. Faddeeva, Computation Methods of Linear Algebra, W.H. Freeman, & Co., San Fransicso, Calif., 1963.

[3] R.W. Bass and I. Gura, "High Order System Design Via State-Space Considerations", Proceedings of the 1965 Joint Automatic Control Conference, Rensselaer Polytechnic Institute, Troy, N.Y., 1965, page 311.

[4] L.A. Zadeh and C.A. Desoer, Linear System Theory: The State Space Approach, McGraw-Hill Book Co., Inc., New York, N.Y., 1963.

[5] C.G.J. Jacobi, Crelle's Journal fur die Reine und Angewandte Mathematik, Vol 30, 1846, pages 51-95.

[6] R. Bellman, Introduction to Matrix Analysis, McGraw-Hill Book Company, Inc., New York, N.Y., 1960.

STABILITY OF SINGLE-LOOP NONLINEAR FEEDBACK SYSTEMS†

CHUNG T. LEE and CHARLES A. DESOER
Department of Electrical Engineering
University of California
Berkeley, California

ABSTRACT

The object of this paper is to find sufficient conditions of stability of the feedback systems shown in Fig. 1.

The subsystem N is a memoryless nonlinearity having its graph in a given sector. The subsystem K is a dynamical system whose input-output relation is described by an operator K as follows:

$$y(t) = z(t) + (K\alpha)(t) \qquad t \geq 0$$

where $z(\cdot)$ may represent a disturbance at the output or a zero-input response of the subsystem K.

The sufficient conditions obtained reduce to Popov's stability conditions if the operator K is assumed to be a convolution operator.

Since no assumptions are made upon the operator K about linearity or time-invariance, the analysis is carried out in a function space.

The case when the operator K is linear is discussed and extension to higher order system is included.

1. INTRODUCTION

We consider the system S shown in Fig. 1. It is a single-input single-output feedback system. Popov[1] obtained sufficient conditions for stability with the assumption that K is a stable time-invariant linear system which is characterized by a differential equation. Recently Ibrahim and Rekasius[2] extended Popov's results to a special class of higher order systems, and Desoer[3] considered the case when the input-output relation of K is described by a convolution integral.

In this paper we shall study the system S without making assumptions of linearity and time-invarance on K and we obtain some sufficient conditions for stability.

Further generalizations can be found in Ref. 4.

2. DEFINITIONS AND ASSUMPTIONS

We shall introduce some definitions and notations used throughout this paper.

Definitions

1. L^2: a set of real valued functions which are square integrable over $(0, \infty)$; $< \cdot, \cdot >$ will denote scalar product in L^2, $<x, y> \triangleq \int_0^\infty x(t)\, y(t)\, dt$, and $\|\cdot\|$ the corresponding norm.
2. L_e^2: a set of real valued functions which are square integrable over $(0, T)$, for all finite $T > 0$.

---

† The research reported herein was supported by National Aeronautics and Space Administration under Grant NsG-354 (S-1).

3. L^{∞} : a set of real valued bounded functions over $(0, \infty)$.

4. Let x be a function in L_e^2. Let us call P_T the projection such that $(P_T x)(t) = x(t)$ for $0 \leq t \leq T$, $(P_T x)(t) = 0$ elsewhere. We shall often write x_T for $P_T x$ and we adopt the following conventions

$$\|x\|_T \triangleq \|x_T\| \qquad \forall T \geq 0 \qquad \forall x \in L_e^2$$

$$< x, y >_T \triangleq < x_T, y_T > \qquad \forall T \geq 0 \qquad \forall x, y \in L_e^2$$

5. An operator mapping from L_e^2 into itself is said to be nonanticipative if and only if

$$P_T(Kx) = P_T\{K(P_T x)\} \qquad \forall x \in L_e^2, \qquad \forall T \geq 0$$

6. The operator D is the time derivative operator, i.e.

$$(Dx)(t) \triangleq \frac{dx(t)}{dt}$$

7. The system S is said to be zero-input L^2-stable (zero-input L^{∞}-stable) if the zero-input response y of the system S is square integrable (bounded) over $(0, \infty)$ for all initial states.

8. The system S is said to be asymptotically stable in the large if it is zero-input L^{∞}-stable and the zero-input response goes to zero as $t \longrightarrow \infty$.

Assumptions

We shall make some assumptions on the subsystems N and K.

N1. The subsystem, the block labeled N, is characterized by

$$\alpha(t) = \phi[e(t)] \tag{1}$$

N2. There is a positive constant k such that $0 \leq e\phi(e) \leq ke^2$ for $-\infty < e < \infty$, and $\phi(0) = 0$.

N3. $\int_0^{\sigma} \phi(\xi)\, d\xi \longrightarrow \infty$ as $|\sigma| \longrightarrow \infty$

K1. The subsystem K is characterized by its input-output relation

$$y = z + K\alpha \tag{2}$$

where $z(\cdot)$ is the zero-input response which depends on the state of K at time 0 and K is a nonanticipative operator mapping from L^2 into itself. Here K has been used to label the subsystem as well as the operator, but the possible ambiguity will be removed by the context.

K2. The operator DK is also assumed to be a nonanticipative operator mapping from L_e^2 into itself where

$$(DKx)(t) = \frac{d}{dt}(Kx)(t).$$

K3. For all initial states, $z(\cdot)$ and $\dot{z}(\cdot)$ are in L^2 and $y(0)$ is a finite number.

3. STABILITY OF THE SYSTEM

Of course since K is not necessarily linear or time-invariant there is no frequency domain interpretation. However, it is of interest to note that the subsystem K must satisfy some simple conditions.

In the preceding section, N is to have its graph in the sector bounded by $\alpha = 0$ and $\alpha = ke$. K1 requires that the zero-input response of K appears as an additive term in (2): this is a strong restriction on K since K is not necessarily linear. In the following section, we assume that K is linear, then the form of (2) is a natural consequence of linearity. Also we may interpret z as the response due not only to initial states of K but also to an additive disturbance at the output of the subsystem K. K3 with K1 assumes that the open loop system is L^2-stable which is a natural assumption because as a special case of N we may consider $\phi(e) \equiv 0$ for all e. The assumption K3 with K2 guarantees that $y(\cdot)$ is differentiable and hence in L_e^2.

The main results are stated in the form of six theorems.

Theorem1

Consider the system S shown in Fig. 1. Let $u \equiv 0$. Suppose that the assumptions N1-N3 and K1-K3 are satisfied. Under these assumptions if there are positive numbers q and δ such that

$$< x, \ (K + qDK + \frac{1}{k}) x > \ \geq \delta \ \|x\|^2 \text{ for all } x \in L^2 \tag{3}$$

then the system S is zero-input L^∞-stable and zero-input L^2-stable.

Remark1

The hypothesis of (3) is nothing but the condition given by Desoer[3] if K is assumed to be linear time-invariant.

Here q does not have to be a constant. Further generalization may be found in Ref. 4.

Before proving the theorem above we state an auxilary lemma. The proof of this lemma is given in Appendix 1.

Lemma 1

Let A be an operator mapping a Hilbert space H into itself, and let x_i, i=1, 2, 3 be in H such that

$$x_1 = x_2 + Ax_3$$

Under the above assumptions if there is a real positive number δ such that

$$< x, \ Ax > \ \geq \delta \ \|x\|^2 \qquad \forall x \in H.$$

then

$$- < x_1, \ x_3 > \ \leq \frac{1}{4\delta} \ \|x_2\|^2 .$$

Here $< \cdot , \cdot >$ and $\|\cdot\|$ denote scalar product and norm in H respectively.

Now we proceed to the proof of Theorem1 based on Lemma1.

Proof of Theorem1. Since K and DK are nonanticipative operators mapping L_e^2 into itself

$$P_T (K\alpha) = P_T \{ K(P_T \alpha) \}$$

and $$\forall T \geq 0, \quad \forall \alpha \in L_e^2$$

$$P_T (DK\alpha) = P_T \{ DK (P_T\alpha) \}$$

Note that y, $\dot{y}$, and α are in L_e^2.

Differentiating (2) we get

$$\dot{y} = \dot{z} + DK\alpha \cdot$$

Let μ be a constant satisfying $0 < \mu < \delta$, then

$$y_T + q\dot{y}_T + (\frac{1}{k} - \mu)\alpha_T = z_T + q\dot{z}_T + P_T\{(K + qDK + \frac{1}{k} - \mu)\alpha_T\} \forall T \geq 0$$

Note that (3) implies that

$$< x, (K + qDK + \frac{1}{k} - \mu)x > \geq (\delta - \mu)\|x\|^2 \quad \forall x \in L^2$$

where $\delta - \mu > 0$. Hence by Lemma 1 for any finite T

$$-\int_0^T \{y(t) + q\dot{y}(t) + (\frac{1}{k} - \mu)\alpha(t)\}\alpha(t)\,dt \leq \frac{1}{4(\delta - \mu)}\|z + q\dot{z}\|_T^2$$

Since $z + q\dot{z} \in L^2$, $u \equiv 0$ implies $e = -y$, and $\alpha = \phi(e)$, we have

$$\int_0^T \{e(t) - (\frac{1}{k} - \mu)\phi[e(t)]\}\phi[e(t)]\,dt + q\int_0^T \phi[e(t)]\,\dot{e}(t)\,dt \leq C \tag{4}$$

where

$$C \triangleq \frac{1}{4(\delta - \mu)}\|z + q\dot{z}\|^2.$$

Observe that the constant C is a finite number independent of T. Call J_1 and J_2 the two terms of the left hand side of (4)

$$J_1 = \int_0^T \{e(t) - \frac{1}{k}\phi[e(t)]\}dt + \mu\int_0^T \phi^2[e(t)]\,dt$$

Because N2 implies $\{e - \frac{1}{k}\phi[e]\}\phi[e] \geq 0$,

$$J_1 \geq \mu\int_0^T \phi^2[e(t)]\,dt \geq 0 \tag{5}$$

Therefore

$$J_2 = q\int_0^T \phi[e(t)]\,\dot{e}(t)\,dt = q\int_0^{e(T)} \phi(\xi)\,d\xi - q\int_0^{e(0)} \phi(\xi)\,d\xi \leq C$$

$$q\int_0^{e(T)} \phi(\xi)\,d\xi \leq C' \quad \text{where} \quad C' \triangleq C + q\int_0^{e(0)} \phi(\xi)\,d\xi \tag{6}$$

Since $y(0)$ is a finite number by K3 and $e(0) = -y(0)$, the constant C' is a finite number independent of T. With N2 and N3, Eq. (6) implies

$$|e(T)| \leq C'' \quad \forall T \geq 0$$

where C'' depends only on $\frac{1}{q}C'$.

Therefore $e \in L^\infty$ and consequently $y \in L^\infty$. Now observe that $J_1 \leq C'$

With Eq (5) $\quad \phi[e] \in L^2$

Because $K : L^2 \longrightarrow L^2$ and $z(\cdot) \in L^2$, $y(\cdot) \in L^2$

Therefore

$$y(\cdot) \in L^2 \cap L^\infty$$

Hence the system S is zero-input L^∞-state and zero-input L^2-stable.

Q. E. D.

Remark 2

In the proof of Theorem 1, N3 is not required to show that S is zero-input L^2-stable.

Remark 3

If $\sup_{\substack{\alpha \in L^2 \\ \alpha \neq 0}} \frac{\|DK\alpha\|}{\|\alpha\|} < \infty$, then Theorem 1 holds for $q = 0$. In other words if $\sup_{\substack{\alpha \in L^2 \\ \alpha \neq 0}} \frac{\|DK\alpha\|}{\|\alpha\|} < \infty$ and $<x, (K + \frac{1}{k})\, x> \;\geq \delta \|x\|^2$ for all $x \in L^2$, then we can find two positive numbers q, δ' such that

$$< x, (K + qDK + \frac{1}{k})\, x > \;\geq \delta' \|x\|^2 \quad \forall x \in L^2$$

i.e. let q be sufficiently small such that

$$\delta' \triangleq \delta - q \sup \frac{\|DK\alpha\|}{\|\alpha\|} > 0$$

Theorem 2

If in addition to all hypotheses in Theorem 1 we have $DK: L^\infty \to L^\infty$ (or $L^2 \to L^\infty$) and $\dot{z}$ is bounded, then the system S is asymptotically stable in the large.

Proof

By Theorem 1 we have $y(\cdot) \in L^\infty \cap L^2$ and $\alpha(\cdot) \in L^\infty \cap L^2$ and $\dot{y} = \dot{z} + DK\alpha$. If either $DK: L^\infty \to L^\infty$ or $DK: L^2 \to L^\infty$, then $DK\alpha$ is bounded. Since $\dot{z}$ is also bounded by assumption, $\dot{y}$ is bounded. Hence $y(t) \to 0$ as $t \to \infty$.

Q.E.D.

Theorems 1 and 2 provide us sufficient conditions for L^∞-stability, L^2-stability, and asymptotic stability. The significance of this derivation is the fact that these conditions are derived without using Liapunov's functions.

It might seem that the conclusions are restricted to statements concerning the zero input response of the closed loop systems, but this is not the case; suppose that $u \neq 0$ and that it satisfies the conditions satisfied by z. Then since (2) may be rewritten as $-e = z - u + K\alpha$ with $e = u - y$ and since $z - u$ in the present case satisfies the same conditions as z in the statement of the theorems, the same conclusions hold with inputs belonging to this class.

4. LINEAR SUBSYSTEMS

In this section we shall assume that the subsystem K is linear time-varying which will be denoted by G to distinguish from the general nonlinear subsystem K considered in the preceding section. The system under consideration is shown in Fig. 2.

We shall investigate the stability of this class of systems. We first state some assumptions on G.

G1. The subsystem G is characterized by a superposition integral

$$y(t) = z(t) + \int_0^t g(t, \tau)\, \alpha(\tau)\, d\tau \qquad t \geq 0 \tag{7}$$

or in an operator equation,

$$y = z + G\alpha$$

and

$$g(t, \tau) = 0 \qquad t < \tau .$$

G2.

$$\int_0^\infty | g(t_1, t_2) |\, dt_1 \leq M_1 < \infty \qquad \forall t_2$$

$$\int_0^\infty | g(t_1, t_2) |\, dt_2 \leq M_2 < \infty \qquad \forall t_1$$

$g(t, t)$ is continuous.

G3. $g(t_1, t_2)$ is differentiable with respect to t_1 and

$$\int_0^T \left| \frac{\partial}{\partial t_1} g(t_1, t_2) \right| dt_i < \infty \text{ for all } T \geq 0 \qquad (i = 1, 2)$$

G4. $z(\cdot)$ and $\dot{z}(\cdot)$ are in L^2 and $z(0)$ is finite.

Theorem 3

For the system shown in Fig. 2 suppose that N1-N3 and G1-G4 are satisfied. Under these assumptions, if there are two positive constants q, δ such that

$$< x, (G + qDG + \tfrac{1}{k})\, x > \geq \delta \, \| x \|^2 \qquad \forall x \in L^2$$

then the system is zero-input L^∞-stable and zero-input L^2-stable.

Proof

It suffices to prove that G1-G4 imply that K1-K3.

Since $g(t, \tau) = 0$ $t < \tau$, the operators G and DG are nonanticipative.

1) $G : L^2 \longrightarrow L^2$ by G2; indeed

$$\| G\alpha \|^2 \leq \int_0^\infty \left\{ \int_0^\infty | g(t, \tau) |\, |\alpha(\tau)|\, d\tau \right\}^2 dt \leq M_1 M_2 \, \|\alpha\|^2 .$$

2) $DG : L^2_e \longrightarrow L^2_e$

$$(DGx)(t) = \int_0^t \frac{\partial}{\partial t} g(t, \tau)\, x(\tau)\, dt + g(t, t)\, x(t).$$

G3 and Schwarz's inequality imply that, for all $x \in L^2_e$, the first term is in L^2_e; since $g(t, t)$ is continuous, $g(t, t)\, x(t)$ is in L^2_e for all $x \in L^2_e$.

Further assumptions must be made to establish asymptotic stability. There are many ways to obtain asymptotic stability with some additional assumptions on G. Here we shall consider two of such cases.

Theorem 4

If in addition to all assumptions in Theorem 3

$$\int_0^\infty \left| \frac{\partial}{\partial t_1} g(t_1, t_2) \right| dt_i \leq N_i < \infty \qquad i = 1, 2,$$

where N_1 is independent of t_2 and N_2 is independent of t_1 and $g(t, t)$ is bounded and $\dot{z}(\cdot)$ is bounded, then the system is asymptotically stable in the large.

Proof

The additional assumption guarantees that $DG: L^\infty \longrightarrow L^\infty$. Hence by Theorem 2 the system is asymptotically stable in the large.

Theorem 5

If in addition to all the assumptions in Theorem 3 we have

$z(t) \longrightarrow 0 \quad \text{as} \quad t \longrightarrow \infty$

$|g(t_1, t_2)| \leq g_M < \infty$

$g(t_1, t_2) \longrightarrow 0$ as $t_1 \longrightarrow \infty$ uniformly in t_2, for t_2 in any finite interval, then the system is asymptotically stable in the large.

This theorem is a direct consequence of the following lemma. The proof of this lemma shall be given in Appendix 2.

Lemma 2

Let x be in $L^\infty \cap L^2$. Suppose that

1. $g(t_1, t_2) = 0 \qquad t_1 < t_2$
2. $|g(t_1, t_2)| \leq g_M < \infty \qquad \forall t_1, t_2$
3. $\int_0^\infty |g(t_1, t_2)| \, dt_2 \leq M_2 < \infty, \qquad \forall t_1$
4. $g(t_1, t_2) \longrightarrow 0$ as $t_1 \longrightarrow \infty$ uniformly in t_2, for t_2 in any finite interval.

Then

$$w(t) \triangleq \int_0^t g(t, \tau) \, x(\tau) \, d\tau. \text{ goes to zero as } t \longrightarrow \infty.$$

Example: It can easily be verified that with $g(t, \tau) = e^{-(2t - \tau)} \sin(t - \tau) \, u(t - \tau)$ where u(t) is a unit step function,

$$< x, \left(G + \frac{1}{2} DG + \frac{1}{k}\right) x > \geq \frac{1}{k} \| x \|^2$$

5. EXTENSION TO HIGHER ORDER SYSTEMS

All the results obtained for single-input single-output feedback systems can be extended to the case when we have multiple-input multiple-output feedback systems. The block diagram representation is shown in Fig. 3.

Assumptions corresponding to higher order systems shall be stated: $\mathcal{N}$ 1. The subsystem, $\mathcal{N}$, is characterized by

$$\underline{\alpha}(t) = \underline{\phi}[\underline{e}(t)]$$

where $\phi(\cdot)$ is a piecewise continuous function mapping $E^n \longrightarrow E^n$
$\mathcal{N}$ 2. There is a constant $n \times n$ matrix $\underline{M}$ such that

$$\underline{\phi}'(\underline{e})\ \underline{e} \geq \underline{\phi}'(\underline{e})\ \underline{M}\ \underline{\phi}(\underline{e})$$

and

$$\underline{\phi}(\underline{0}) = \underline{0}$$

Here the prime denotes transposition.

$\mathcal{N}$ 3. Let $V: E^n \longrightarrow E$ and let $\underline{Q}$ be a $n \times n$ constant symmetric matrix such that

$$\underline{Q}\ \underline{\phi}(\underline{e}) = \nabla V(\underline{e}) \qquad \forall e \in E^n$$

and

$$V(\underline{e}) \longrightarrow +\infty \quad \text{as} \quad |\underline{e}| \longrightarrow \infty$$

Here

$\nabla V(\underline{e})$ is the gradient of $V(\underline{e})$.

$\mathcal{K}$ 1. $\quad \underline{y} = \underline{x} + \underline{K}\,\underline{\alpha}$

where $\underline{K}$ is a nonanticipative operator mapping L_n^2 into itself. Here L_n^2 is

$$L_n^2 \triangleq \{\underline{x}(\cdot) \mid \int_0^\infty \underline{x}'(t)\ \underline{x}(t)\ dt < \infty\}$$

$\mathcal{K}$ 2. The operator $D\,K$ is also a nonanticipative operator mapping L_{ne}^2 into itself.

$$L_{ne}^2 \triangleq \{\underline{x}(\cdot) \mid \int_0^T \underline{x}'(t)\ \underline{x}(t)\ dt < \infty \text{ for all finite } T\}$$

$\mathcal{K}$ 3. $\underline{z}(\cdot)$ and $\dot{\underline{z}}(\cdot)$ are in L_n^2 and $|\underline{y}(0)|$ is a finite number.

A typical extension shall be given in the following theorem.

Theorem 6

Suppose that $\mathcal{N}$ 1-$\mathcal{N}$3 and $\mathcal{K}$ 1-$\mathcal{K}$ 3 are satisfied. Under these assumptions if there is a positive number δ such that

$$< \underline{x}, (\underline{K} + \underline{Q}\,D\,\underline{K} + \underline{M})\ \underline{x} > \ \geq \delta\, \|\underline{x}\|^2, \quad \forall \underline{x} \in L_n^2$$

then the zero-input response of closed loop system is bounded and in L_n^2.

Proof

Proof is almost identical to that of Theorem 1. The following equality is necessary

$$\langle \underline{e}, \nabla V(\underline{e}) \rangle_T = V(e(T)) - V(e(0)).$$

REFERENCES

1. M. Aizerman and F. R. Gantmacher, Absolute Stability of Regulator Systems, (Translated from Russian by E. Polak). San Francisco; Holden Day, 1964.
2. E. S. Ibrahim and Z. V. Rekasius, "A Stability Criterion for Nonlinear Feedback Systems," IEEE Trans. on Automatic Control Vol. AC-9, No. 2, p. 154-159, April 1964.
3. C. A. Desoer, "A Generalization of the Popov Criterion," IEEE Trans. on Automatic Control, Vol. AC-10, No. 2, p. 182-184, April 1965.
4. C. T. Lee, "Study of Stability of Nonlinear Single Loop Feedback Systems," Ph.D. Thesis, Dept. of Electrical Engineering, Univ. of California, Berkeley, 1966.

APPENDIX 1

Proof of Lemma 1

$$\langle Ax, x \rangle \geq \delta \, \| x \|^2 \quad \forall x \in H$$

$$x_1 = x_2 + A x_3$$

$$\langle x_1, x_3 \rangle = \langle x_2, x_3 \rangle + \langle x_3, A x_3 \rangle \geq \langle x_2, x_3 \rangle + \delta \langle x_3, x_3 \rangle$$

$$= \left\| \sqrt{\delta} \; x_3 + \frac{1}{2\sqrt{\delta}} x_2 \right\|^2 - \frac{1}{4\delta} \| x_2 \|^2 \geq - \frac{1}{4\delta} \| x_2 \|^2$$

$$\therefore \; -\langle x_1, x_3 \rangle \leq \frac{1}{4\delta} \| x_2 \|^2$$

Q.E.D.

APPENDIX 2

Proof of Lemma 2

$$| w(t) |^2 = \left\{ \int_0^t g(t, \tau) \, x(\tau) \, d\tau \right\}^2 \leq \left\{ \int_0^t | g(t, \tau) | \; | x(\tau) | \, d\tau \right\}^2$$

$$\leq M_2 \int_0^t | g(t, \tau) | \, (x(\tau))^2 \, d\tau$$

Since $x \in L^2$, for arbitrary $\epsilon_1 > 0$ we can choose T_1 sufficiently large so that

$$\int_{T_1}^{\infty} (x(\tau))^2 \, d\tau < \epsilon_1 .$$

By assumption (4) of Lemma 2, given any $\epsilon_2 > 0$ and any $T_1 > 0$ there is a $T_2(\epsilon_2, T_1)$ such that

$$\int_0^{T_1} |g(t,\tau)|\,d\tau < \epsilon_2 \text{ for } t \geq T_2$$

Let $\sup\limits_{0 \leq t < \infty} |x(t)| = X_M$

Then

$$\int_0^t |g(t,\tau)|\,|x(\tau)|^2\,d\tau = \int_0^{T_1} |g(t,\tau)|\,|x(\tau)|^2\,d\tau$$

$$+ \int_{T_1}^t |g(t,\tau)|\,|x(\tau)|^2\,d\tau$$

$$\leq X_M^2 \int_0^{T_1} |g(t,\tau)|\,d\tau + g_M \int_{T_1}^T |x(\tau)|^2\,d\tau$$

$$\leq X_M^2 \int_0^{T_1} |g(t,\tau)|\,d\tau + g_M \int_{T_1}^\infty |x(\tau)|^2\,d\tau$$

Hence,

$$\int_0^t |g(t,\tau)|\,|x(\tau)|^2\,d\tau \leq X_M^2\epsilon_2 + g_M\epsilon_1 \quad \text{for } t \geq T_2 .$$

$$\therefore\ |w(t)|^2 \leq M_2 X_M^2 \epsilon_2 + g_M \epsilon_1 \quad \text{for } t \geq T_2$$

$$\therefore\ w(t) \longrightarrow 0 \quad \text{as} \quad t \longrightarrow \infty .$$

Q.E.D.

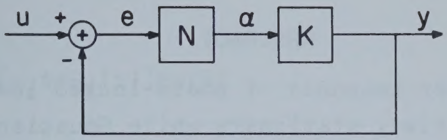


Fig. 1. The System S.

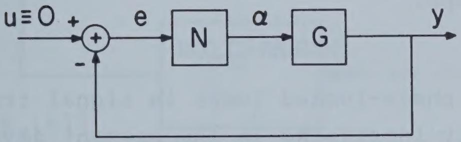


Fig. 2. Linear Subsystem

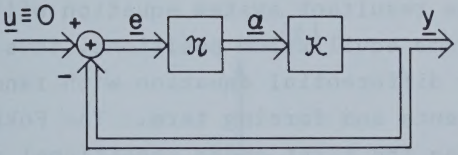


Fig. 3. Higher Order System

STABILITY CRITERIA OF PHASE-LOCKED LOOPS

P. W. U. GRAEFE* and N. K. LOH**
*National Research Council of Canada, Ottawa, Ontario.
**University of Waterloo, Waterloo, Ontario, Canada.

ABSTRACT

The nonlinear response of phase-locked loops in the presence of additive, stationary white Gaussian noise is linearized about some equilibrium point. The Fokker-Planck equation is set up for the linearized model, from which the moments of the response are obtained. Criteria for stability in mean and mean-square of the linearized system are then established. Results are obtained both for first-order and higher order loops.

INTRODUCTION

The use of phase-locked loops in signal tracking devices is steadily increasing in the present days of rapidly advancing space technology. Due to the inherent nonlinearity the phase-locked loop is also difficult to analyse.

In this paper we will first obtain a nonlinear model of the phase-locked loop which is subject to an additive noise back-ground. The resultant system equation will then be linearised about some equilibrium positions, thus leading to a linear nth order differential equation with random, time varying coefficients and forcing term. The Fokker-Planck equation governing the first order conditional probability density of the linearised system will then be obtained, and the differential equation governing the nth moment of the system response derived from it. Conditions for stability in mean and mean square as well as for higher order moments will then be established and the threshold values of noise intensity, beyond which it is impossible for the phase-locked loop to operate, will be obtained.

DERIVATION OF THE NONLINEAR MODEL

The block diagram for the phase-locked loop is as shown in Fig. 1. The input signal $v_i(t)$ of the multiplier is assumed to have passed through a bandpass filter with center frequency ω_c. The noise n(t) is assumed to be a stationary wide band Gaussian process with spectral density as shown in Fig. 2. For practical purposes this noise may be assumed to be white.

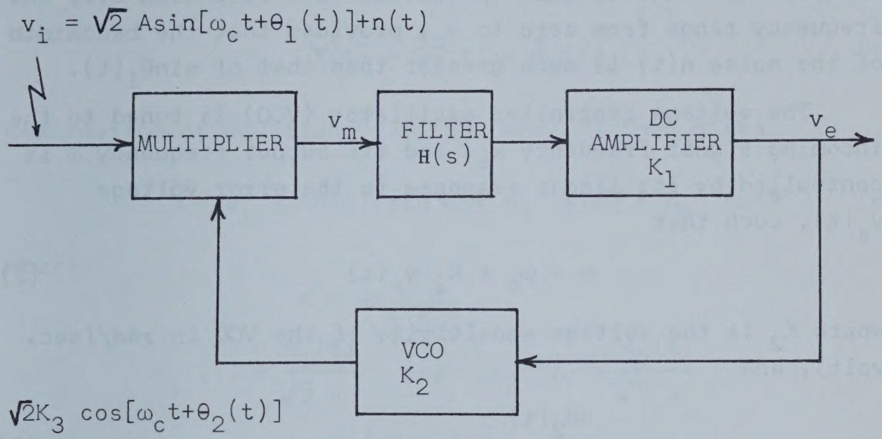

$v_i = \sqrt{2}\,A\sin[\omega_c t+\theta_1(t)]+n(t)$
MULTIPLIER
v_m
FILTER
H(s)
DC
AMPLIFIER
K_1
v_e
VCO
K_2
$\sqrt{2}K_3\cos[\omega_c t+\theta_2(t)]$

FIGURE 1: Phase-Locked Loop

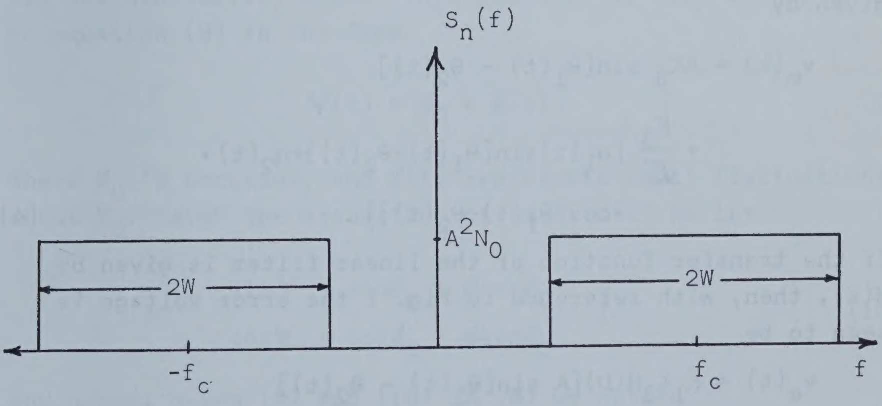


FIGURE 2: Spectrum of n(t)

Let us express the noise n(t) in terms of its orthogonal components $n_1(t)$, $n_2(t)$:

$$n(t) = n_1(t)\sin[\omega_c t + \theta_1(t)] + n_2(t)\cos[\omega_c t + \theta_1(t)] \quad (1)$$

where $n_1(t)$ and $n_2(t)$ can be shown (Van Trees, 1963) to be independent white Gaussian processes with zero mean over the frequency range from zero to ω_c, provided that the bandwidth of the noise n(t) is much greater than that of $\sin\theta_1(t)$.

The voltage controlled oscillator (VCO) is tuned to the incoming signal frequency ω_c, and its output frequency ω is controlled by its linear response to the error voltage $v_e(t)$, such that

$$\omega = \omega_c + K_2 v_e(t) \quad (2)$$

where K_2 is the voltage sensitivity of the VCO in rad/(sec. volt), and

$$\frac{d\theta_2(t)}{dt} = K_2 v_e(t) . \quad (3)$$

With reference to Fig. 1 and noting that both the filter and the VCO will not respond to the double frequency terms, we see that the multiplier output which is of interest is given by

$$\begin{aligned} v_m(t) = {} & AK_3 \sin[\theta_1(t) - \theta_2(t)] \\ & + \frac{K_3}{\sqrt{2}}\{n_1(t)\sin[\theta_1(t)-\theta_2(t)]+n_2(t)\cdot \\ & \cdot\cos[\theta_1(t)-\theta_2(t)]\} . \end{aligned} \quad (4)$$

If the transfer function of the linear filter is given by H(s), then, with reference to Fig. 1 the error voltage is seen to be

$$\begin{aligned} v_e(t) = {} & K_1K_3H(D)\{A\sin[\theta_1(t) - \theta_2(t)] \\ & + \frac{n_1(t)}{\sqrt{2}}\sin[\theta_1(t) - \theta_2(t)] + \frac{n_2(t)}{\sqrt{2}} \cdot \\ & \cdot\cos[\theta_1(t) - \theta_2(t)]\} \end{aligned} \quad (5)$$

where H(D), (D = d/dt), represents the operator of the filter in the time domain, operating on the filter input.

If we let $\Psi(t)$ denote the phase difference between the received signal and the VCO output, that is

$$\Psi(t) = \theta_1(t) - \theta_2(t) \tag{6}$$

then

$$\dot{\Psi}(t) = \dot{\theta}_1(t) - \dot{\theta}_2(t). \tag{7}$$

Combination of equations (3), (5) and (7) then gives

$$\dot{\Psi} = \dot{\theta}_1 - KH(D)(\sin\Psi + n_1' \sin\Psi + n_2' \cos\Psi) \tag{8}$$

where

$$K = AK_1K_2K_3 ,$$

$$n_1'(t) = \frac{n_1(t)}{\sqrt{2}\,A} \quad \text{and} \quad n_2'(t) = \frac{n_2(t)}{\sqrt{2}\,A} .$$

LINEARISATION OF THE SYSTEM EQUATION

Examination of phase-locked loops under the influence of input noise indicates that the phase difference $\Psi(t)$ exhibits some random variation about some mean positions which are not necessarily zero. This leads us to seek a solution to equation (8) in the form

$$\Psi(t) = \phi_0 + \phi(t) \tag{9}$$

where ϕ_0 is constant, and $\phi(t)$ represents small fluctuations about ϕ_0. With the assumption of small $\phi(t)$ we let

$$\left.\begin{aligned} \sin\Psi &= \sin\phi_0 + \phi\cos\phi_0 \\ \cos\Psi &= \cos\phi_0 - \phi\sin\phi_0 \end{aligned}\right\} \tag{10}$$

and hence, using (9) and (10) in (8) we obtain

$$\dot{\phi} = \dot{\theta}_1 - KH(D)\{[m_2(t)+\cos\phi_0]\phi + m_1(t)+\sin\phi_0\} \tag{11}$$

where

$$\left.\begin{aligned} m_1(t) &= n_1'(t)\sin\phi_0 + n_2'(t)\cos\phi_0 \\ m_2(t) &= n_1'(t)\cos\phi_0 - n_2'(t)\sin\phi_0 \end{aligned}\right\} \qquad (12)$$

It is easily seen from (12) that $m_1(t)$ and $m_2(t)$ are independent white Gaussian processes (in the interval from zero to ω_c) with zero mean and identical correlation functions. That is

$$\left.\begin{aligned} &\langle m_1(t)\rangle = 0 = \langle m_2(t)\rangle \\ &\langle m_i(t_1)m_j(t_2)\rangle = \frac{N_0}{2A^2}\,\delta(t_1-t_2) \qquad i = j = 1,2 \\ &\qquad\qquad\qquad = 0 \qquad\qquad\qquad\quad i \neq j \end{aligned}\right\} \qquad (13)$$

FIRST ORDER SYSTEM

For a filter with unit transfer function, that is $H(s) = 1$, equation (11) represents a first order system. If we let $\theta_1(t) = \omega_0 t$, where ω_0 is the difference between the input and the VCO angular frequency, we obtain from (11) a first order linear differential equation with a coefficient and forcing term which are Gaussian white noise processes, namely:

$$\dot{\phi} = \omega_0 - K\{[m_2(t)+\cos\phi_0]\phi + m_1(t) + \sin\phi_0\}. \qquad (14)$$

It has been shown by Ariaratnam (1960) and Wang and Uhlenbeck (1945) that under the assumption of Gaussian white noise input and coefficient variation the system response $\phi(t)$ of (14) forms a Markov process. For such a process, if $p(\phi,t|\phi',t')d\phi$ denotes the probability that the 'state' of the system lies in the differential element $(\phi,\phi+d\phi)$ of state space at time t, given that it was at ϕ' at time t', then the conditional probability density function $p = p(\phi,t|\phi',t')$ is given by the positive normalized solution of the Fokker-Planck equation

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial\phi}(Ap) + \frac{1}{2}\frac{\partial^2}{\partial\phi^2}(Bp) \qquad (15)$$

satisfying the initial condition:

$$p(\phi,t|\phi',t') \to \delta(\phi-\phi') \quad \text{as } t \to t'^{+} .$$

The incremental moments A and B are given by

$$\left.\begin{aligned} A &= \lim_{\delta t \to 0} \frac{\langle \delta\phi \rangle}{\delta t} \\ B &= \lim_{\delta t \to 0} \frac{\langle (\delta\phi)^2 \rangle}{\delta t} \end{aligned}\right\} \qquad (16)$$

where the ensemble averaging, denoted by $\langle \cdot \rangle$, is performed at some time in the interval δt, assuming conditions at the beginning of the time interval being known. The general method of derivation of these incremental moments are given by Ariaratnam and Graefe (1965). In this case it is found that

$$\left.\begin{aligned} A &= \left(\frac{K^2N_0}{4A^2} - K\cos\phi_0\right)\phi(t) + \omega_0 - K\sin\theta_0 \\ B &= \frac{K^2N_0}{2A^2}\,\phi^2(t) + \frac{K^2N_0}{2A^2} \ . \end{aligned}\right\} \qquad (17)$$

No general method for obtaining explicit solutions of the Fokker-Planck equation has yet been found. For a stability analysis, however, we find that we only require differential equations governing the moments of the response, and these are quite easily derived from the Fokker-Planck equation as follows.

To obtain the differential equation for the Nth moment, we multiply both sides of (15) by ϕ^N and integrate by parts over the entire space of ϕ.
Thus

$$\left.\begin{aligned} &\frac{d}{dt}\langle \phi^N \rangle + N\left(K\cos\phi_0 - N\frac{K^2N_0}{4A^2}\right)\langle \phi^N \rangle \\ &\qquad = N(\omega_0 - K\sin\phi_0)\langle \phi^{N-1} \rangle + N(N-1)\frac{K^2N_0}{4A^2}\langle \phi^{N-2} \rangle \ . \end{aligned}\right\} \qquad (18)$$

We see from (18) that the Nth moment only depends on moments lower than N, and hence the moments may be evaluated recursively, starting from N = 1. We also see that the moments of order N are bounded if

(a) the lower order moments are bounded, and

(b)
$$\cos\phi_0 > N \frac{KN_0}{4A^2} . \tag{19}$$

If we knew the equilibrium value ϕ_0, we would have the necessary and sufficient conditions for Nth order moment stability. With ϕ_0 unknown, however, we still have $-1 \leq \cos\phi_0 \leq 1$. Therefore no stable moment of order N can exist if

$$N \frac{KN_0}{4A^2} > 1 ,$$

and hence the condition

$$N \frac{KN_0}{4A^2} < 1$$

is a necessary condition for the boundedness of the Nth moment of the linearized model of the phase-locked loop.

In particular we find from N = 1 and N = 2 respectively that a necessary condition for stability in mean* is that

$$\frac{KN_0}{4A^2} < 1$$

and for stability in mean square is that

$$\frac{KN_0}{2A^2} < 1.$$

---

*An nth order system described by the n state variables y_i, $i = 1,\ldots,n$, is said to be stable in mean if

$$\lim_{t \to \infty} \left| \langle y_i(t) \rangle \right| < K_i \quad (i = 1,2,\ldots,n)$$

and stable in mean square if

$$\lim_{t \to \infty} \left| \langle y_i(t) y_j(t) \rangle \right| < K_{ij} \quad (i,j = 1,2,\ldots,n)$$

where K_i, K_{ij} are positive constants.

HIGHER ORDER LOOPS

Let the transfer function of the filter be given by the ratio of two polynomials in s

$$H(s) = \frac{F(s)}{G(s)} = \frac{\sum_{i=0}^{n-1} c_i s^i}{s^{n-1} + \sum_{j=1}^{n-2} b_j s^j} \tag{20}$$

where deg $F(s) \leq$ deg $G(s)$ and $c_0 \neq 0$.

Also let

$$\emptyset = F(D)X \tag{21}$$

where X is a random variable.

From equations (11), (20) and (21), noting that $\frac{d^i}{dt^i}\theta_1 = 0$, $i \geq 2$ we obtain

$$DG(D)X = -K\{[m_2(t)+\cos\emptyset_0]F(D)X + m_1(t)+\sin\emptyset_0]$$

or

$$\frac{d^n}{dt^n} X + \sum_{r=1}^{n} [a_r+\alpha_r(t)] \frac{d^{r-1}}{dt^{r-1}} X = a_0 + \alpha_0(t) \tag{22}$$

where

$$\left.\begin{aligned}
a_r &= b_{r-2} + Kc_{r-1}\cos\emptyset_0 && (r = 3,4,\ldots,n)\\
a_r &= Kc_{r-1}\cos\emptyset_0 && (r = 1,2)\\
\alpha_r(t) &= Kc_{r-1}m_2(t) && (r = 1,2,\ldots,n)\\
a_0 &= -K\sin\emptyset_0 &&\\
\alpha_0(t) &= -K\, m_1(t)\ . &&\\
\text{Let}\quad X &= y_1,\quad \frac{d^i}{dt^i} X = y_{i+1}, && (i = 0,1,\ldots,n-1)
\end{aligned}\right\} \tag{23}$$

then (22) becomes

$$\frac{d}{dt}\begin{bmatrix} y_1\\ \cdot\\ \cdot\\ \cdot\\ \cdot\\ y_n \end{bmatrix} = -\begin{bmatrix} 0 & -1 & 0 & \ldots & 0\\ 0 & 0 & -1 & & \cdot\\ & & & \ddots & \\ & & & & -1\\ a_1 & \cdot & \cdot & \cdot & a_n \end{bmatrix}\begin{bmatrix} y_1\\ \cdot\\ \\ \cdot\\ y_n \end{bmatrix} - \begin{bmatrix} 0 & \ldots & 0\\ \cdot & & \cdot\\ & & \\ 0 & \ldots & 0\\ \alpha_1 & \ldots & \alpha_n \end{bmatrix}\begin{bmatrix} y_1\\ \cdot\\ \cdot\\ \cdot\\ y_n \end{bmatrix} + \begin{bmatrix} 0\\ \cdot\\ \cdot\\ 0\\ a_0 + \alpha_0 \end{bmatrix} . \tag{24}$$

Again, it has been shown that the response $\underline{y} = (y_1,\ldots,y_n)$ of (24) forms a Markov process since the

$\alpha_i(t)$, $(i = 0,1,\ldots,n)$ are Gaussian white noise processes, and hence, the conditional probability density $p = p(\underline{y},t|y',t')$ satisfies the Fokker-Planck equation

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{n} \frac{\partial}{\partial y_i}(A_i p) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial y_i \partial y_j}(B_{ij}p) \tag{25}$$

where the incremental moments A_i and B_{ij} defined by

$$\left.\begin{aligned} A_i &= \lim_{\delta t \dashrightarrow 0} \frac{\langle \delta y_i \rangle}{\delta t} \\ B_{ij} &= \lim_{\delta t \dashrightarrow 0} \frac{\langle \delta y_i \delta y_j \rangle}{\delta t} \end{aligned}\right\} \tag{26}$$

are found to be (Ariaratnam and Graefe, 1965):

$$\left.\begin{aligned} A_i &= y_{i+1} && (i = 1,2,\ldots,n-1) \\ A_n &= -\sum_{r=1}^{n}(a_r - D_{nr})\, y_r + a_0 \\ B_{ij} &= 0 && (i,j = 1,2,\ldots,n-1) \\ B_{in} &= B_{ni} = 0 && (i = 1,2,\ldots,n-1) \\ B_{nn} &= 2\left[\sum_{r=1}^{n}\sum_{s=1}^{n} D_{rs} y_r y_s + D_{00}\right] \end{aligned}\right\} \tag{27}$$

where

$$\left.\begin{aligned} D_{ij} &= K^2 c_{i-1} c_{j-1} \frac{N_0}{4A^2} && (i,j = 1,2,\ldots,n) \\ D_{00} &= K^2 \frac{N_0}{4A^2} \end{aligned}\right\} \tag{28}$$

Let $m_N(k_1,\ldots,k_n) = \langle y_1^{k_1} \ldots y_n^{k_n} \rangle$ denote the mixed moments of order N, where $N = 1,2,\ldots$ and where $k_1,\ldots,k_n$ are positive integers satisfying $k_1+\ldots+k_n = N$.

Multiplying (25) by $y_1^{k_1} \ldots y_n^{k_n}$ and integrating by parts over the entire state space $-\infty < \underline{y} < \infty$ thus leads to the moment equation

$$\frac{d}{dt} m_N(k_1,k_2,\ldots,k_n) = \sum_{i=1}^{n-1} k_i m_N(k_1,\ldots,k_i-1,k_{i+1}+1, \ldots,k_n) -$$

$$\left.\begin{aligned} &-k_n \sum_{r=1}^{n} (a_r - D_{nr}) m_N(k_1, \ldots, k_r+1, \ldots, k_n-1) \\ &+ a_0 k_n \, m_{N-1}(k_1, \ldots, k_n-1) \\ &+ k_n(k_n-1) D_{00} \, m_{N-2}(k_1, \ldots, k_n-2) \\ &+ k_n(k_n-1) \sum_{r=1}^{n} \sum_{s=1}^{n} D_{rs} \, m_N(k_1, \ldots, k_r+1, \ldots, \\ &\qquad\qquad k_s+1, \ldots, k_n-2) \end{aligned}\right\} \quad (29)$$

We see from (29) that the Nth order moments are related only to moments of lower order and for a particular N there are as many equations as there are mixed moments of that order. Hence, all the higher moments can be calculated recursively once the first moments are found.

The moments of order N are thus bounded if all the moments of order less than N are bounded and the eigenvalues of the coefficient matrix relating the Nth moments to their derivatives have negative real parts.

EXAMPLE OF A SECOND-ORDER LOOP

If we choose H(s) as

$$H(s) = \frac{s+a}{s} \quad (30)$$

we obtain a second order system. From (20), (23) and (28) we see that

$$\left.\begin{aligned} a_0 &= -K\sin\emptyset_0 & D_{00} &= K^2 \frac{N_0}{4A^2} \\ a_1 &= aK\cos\emptyset_0 & D_{11} &= K^2 a^2 \frac{N_0}{4A^2} \\ a_2 &= K\cos\emptyset_0 & D_{22} &= K^2 \frac{N_0}{4A^2} \\ & & D_{12} &= D_{21} = aK^2 \frac{N_0}{4A^2} \end{aligned}\right\} \quad (31)$$

We see from equation (29) that the first order moments, given by:

$$\frac{d}{dt}\begin{bmatrix}\langle y_1\rangle \\ \langle y_2\rangle\end{bmatrix} = \begin{bmatrix}0 & 1 \\ -(a_1-D_{12}) & -(a_2-D_{22})\end{bmatrix}\begin{bmatrix}\langle y_1\rangle \\ \langle y_2\rangle\end{bmatrix} + \begin{bmatrix}0 \\ a_0\end{bmatrix} \qquad (32)$$

are bounded if and only if

$$\cos\emptyset_0 > \frac{KN_0}{4A^2}$$

Since $\langle\emptyset\rangle = a\langle y_1\rangle + \langle y_2\rangle$, it follows that for

$$\frac{KN_0}{4A^2} > 1$$

$\emptyset$ cannot be stable in mean, regardless of the value of $\emptyset_0$. The second moments, given by

$$\left.\begin{aligned}\frac{d}{dt}\begin{bmatrix}\langle y_1{}^2\rangle \\ \langle y_1y_2\rangle \\ \langle y_2{}^2\rangle\end{bmatrix} &= \begin{bmatrix}0 & 2 & 0 \\ -(a_1-D_{12}) & -(a_2-D_{22}) & 1 \\ 2D_{11} & -(2a_1-6D_{12}) & -(2a_2-4D_{22})\end{bmatrix}\begin{bmatrix}\langle y_1{}^2\rangle \\ \langle y_1y_2\rangle \\ \langle y_2{}^2\rangle\end{bmatrix} \\ &\quad + \begin{bmatrix}0 \\ a_0\langle y_1\rangle \\ 2a_0\langle y_2\rangle+2D_{00}\end{bmatrix}\end{aligned}\right\} (33)$$

are bounded if (a) the first moments are bounded

(b) $a_2 - 2D_{22} > 0$, and (34)

(c) $(a_1 - D_{12})(a_2 - 2D_{22}) > D_{11}$.

If this is the case, then

$$\left.\begin{aligned}\lim_{t\to\infty} \langle y_1{}^2\rangle &= D_{00}\Big/[(a_1-D_{12})(a_2-2D_{22})-D_{11}] \\ \lim_{t\to\infty} \langle y_1y_2\rangle &= 0 \\ \lim_{t\to\infty} \langle y_2{}^2\rangle &= D_{00}(a_1-D_{12})\Big/[(a_1-D_{12})(a_2-2D_{22})-D_{11}].\end{aligned}\right\} (35)$$

After the system has been operating for a sufficiently long time such that transients will have died out, we can find the stationary variance σ_{ϕ} of ϕ from (35).

$$\sigma_{\phi} = \langle \phi^2 \rangle = a^2 \langle {y_1}^2 \rangle + a \langle y_1 y_2 \rangle + \langle {y_2}^2 \rangle$$

$$= \frac{N_0 \left(a + K\cos\phi_0 - K^2 \dfrac{N_0}{4A^2} \right)}{4A^2 \cos^2\phi_0 - (3K\cos\phi_0 - 2K^2 + a^2) N_0} \quad . \tag{36}$$

CONCLUSIONS

In this paper a general method has been given for analysing Nth moment stability of an nth order phase-locked loop based on Fokker-Planck techniques. Methods of analysis of stability of phase-locked loops so far given in the literature have limited application in that they are difficult to be applied to either the nth order loops or the investigation of Nth moment stability.

For the case of a first-order loop the results obtained agree with those obtained by Van Trees (1963), except that the method given here seems to be less laborious.

REFERENCES

1. Ariaratnam, S. T., and Graefe, P. W. U., Linear Systems with Stochastic Coefficients II. Accepted for publication, Int. J. Control.

2. Tikhonov, V. I., (a) The Effect of Noise on Phase-Locked Oscillator Operation, Automation and Remote Control, Vol. 20, No. 9, 1959. (b) Phase-Lock Automatic Frequency Control Operation in the Presence of Noise, Automation and Remote Control, Vol. 21, No. 3, 1960.

3. Uhlenbeck, G. E. and Ornstein, L. S., On the Theory of the Brownian Motion, and Wang, M. C. and Uhlenbeck, G. E., On the Theory of the Brownian Motion II, selected papers on Noise and Stochastic Processes, Dover Publication, Inc., New York.

4. Van Trees, H. L., A Lower Bound on Stability in Phase-Locked Loops, Information and Control, Vol.6, No.3, Sept. 1963.

5. Viterbi, A. J., Phase-Locked Loop Dynamics in the Presence of Noise by Fokker-Planck Techniques, Proc. IEEE, Vol. 51, No. 12, December, 1963.

ANALYSIS OF SYSTEMS WITH MULTIPLE CONTINUOUS NONLINEARITIES

STUART A. STEELE
Department of Electrical Engineering
The Pennsylvania State University
University Park, Pennsylvania

ABSTRACT

A method of analysis is proposed and examined for a class of nonlinear systems. It can be applied to systems with one or more nonlinearities. However, it has been developed mainly to investigate systems with multiple nonlinearities and has the inherent property that the number of the nonlinearities is not of major importance. The method involves the roots of a defined characteristic equation associated with the nonlinear differential equation of the system. The roots of this defined characteristic equation are not constant but vary as a function of the nonlinearities. In the class proposed here the differential equation of a system is of order two, the initial value of the dependent variable is known and the first derivative of the dependent variable is zero at t=0. While there are many techniques available for analyzing second order systems, the methods can be quite tedious or not applicable when multiple nonlinearities are involved. This is especially true when many intentional nonlinearities are added to improve system response.

The technique can be applied to systems with continuous nonlinearities in all three terms of the second order differential equation defining the system. A necessary and sufficient condition is developed that can insure whether the response is oscillatory, damped or as fast as possible with no overshoot. The basic method can be used to compare the response of one system to another by developing a performance factor for each system.

Approximations are involved when developing the technique and many examples are given with computer solutions to demonstrate the application of the concepts.

THE EXPONENTIAL DESCRIBING FUNCTION IN THE ANALYSIS OF NONLINEAR SYSTEMS*

THEODORE A. BICKART
Department of Electrical Engineering
Syracuse University
Syracuse, New York 13210

ABSTRACT

Signals in a sub-space, $\mathcal{L}_2(-\infty,t]$, of the space of square integrable signals over the infinite past, $L_2(-\infty,t]$, are approximated by signals in the one dimensional sub-space, $\mathcal{L}_2^1(-\infty,t]$ spanned by the first function from the set of reversed time Laguerre functions. A system mapping $\mathcal{L}_2(-\infty,t]$ into itself is associated with a system mapping $\mathcal{L}_2^1(-\infty,t]$ into itself; the latter system is shown to be a gain — exponential describing function. The exponential describing function is developed as an analysis tool for studying a class of nonlinear feedback systems. The contraction-mapping fixed-point theorem is used to develop conditions for the existence of a solution prior to the use of the exponential describing function to obtain an approximate solution.

INTRODUCTION

The sinusoidal describing function is a valuable tool to the analyst attempting to study the approximate behavior of a nonlinear system subject to periodic excitation. Sinusoidal describing function analysis techniques limit the analyst, however, to the study of periodic, steady state behavior; the study of transient behavior is an entirely different problem.

A similar analysis tool, the exponential describing function, will be developed in this paper. The exponential describing function analysis techniques will provide a means for studying transient behavior of nonlinear systems.

The exponential describing function will be validated as an analysis tool for a class of feedback systems through application of the contraction-mapping fixed-point theorem. The approach to this task will be similar to that taken by Sandberg in justifying the use of the sinusoidal describing function [1] and in validating some truncation techniques in analysis [2].

SIGNAL REPRESENTATION

Let $L_2(-\infty,t]$ denote the space of all real, Lebesgue square integrable functions of the real variable τ defined on $(-\infty,t]$ for $t \epsilon (-\infty, +\infty)$. The signals of interest in this paper will lie in $\mathcal{L}_2(-\infty,t]$, a subspace of $L_2(-\infty,t]$. The time variable, τ, varies over all past time to the present time, denoted by t. $\mathcal{L}_2(-\infty,t]$ will be defined later.

---

*This work was supported by the United States Air Force, Rome Air Development Center under Contract No. AF 30(602)-3538.

Let $\{\varphi_i(\tau)\}$ be a complete, orthonormal set of functions spanning the Hilbert space $L_2(-\infty,t]$ [3,4]. Such a set might be the reversed time Laguerre functions on the semi-infinite interval $(-\infty,t]$. Projection of $q(\tau) \in L_2(-\infty,t]$ onto the set of basis functions $\{\varphi_i(\tau)\}$ yields the coefficients

$$Q_i = \int_{-\infty}^{t} q(\tau)\varphi_i(\tau)d\tau \tag{1}$$

in the series

$$q(\tau) = \sum_{i=1}^{\infty} Q_i \; Q_i(\tau) \tag{2}$$

equation (1) makes it evident that Q_i is a function of the present time t; this will be made manifest on occasion by writing Q_i as $Q_i(t)$. The signals under consideration in this paper may be represented by series such as (2).

Let $L_2^n(-\infty,t]$ be the finite dimensional sub-space of $L_2(-\infty,t]$ having the first n functions from the set $\{\varphi_i(\tau)\}$ as a basis. $\mathcal{L}_2^n(-\infty,t]$ is defined as $\mathcal{L}_2(-\infty,t] \cap L_2^n(-\infty,t]$. In the next section signals in $\mathcal{L}_2(-\infty,t]$ will be approximated by signals in $\mathcal{L}_2^n(-\infty,t]$. In the particular approximation yielding the exponential describing function n will be unity.

EXPONENTIAL DESCRIBING FUNCTION

Let $\mathcal{T}$ denote the set of causal operators which map $\mathcal{L}_2(-\infty,t]$ into itself. The exponential describing function will be developed for systems characterized by operators from the set $\mathcal{T}$.

With $q(\tau) \in \mathcal{L}_2(-\infty,t]$, the projection operator T_P is defined by $T_P q(\tau) = Q_1\varphi_1(\tau)$. The signal $T_P q(\tau)$ is an element of $\mathcal{L}_2^1(-\infty,t]$ and is within some undetermined error, an approximation to $q(\tau)$.

In all subsequent statements, a bar over a symbol for a signal will denote a signal in $\mathcal{L}_2^1(-\infty,t]$ approximating the original signal. That is, if $q(\tau) \in \mathcal{L}_2(-\infty,t]$ then $\bar{q}(\tau) \in \mathcal{L}_2^1(-\infty,t]$ is a signal approximating $q(\tau)$. It will not be necessarily true that $\bar{q}(\tau) = T_P q(\tau)$; this will be obvious.

Now define $\mathcal{T}^1$ to be the set of operators $T_P T$, where T is any member of $\mathcal{T}$. $\mathcal{T}^1 \subset \mathcal{T}$ is obviously a set of operators mapping $\mathcal{L}_2^1(-\infty,t]$ into itself. The operator $T_P T \in \mathcal{T}^1$, corresponding to $T \in \mathcal{T}$, will be written as T^1. If $s(\tau)$ and $r(\tau)$ are the stimulus and response signals, respectively, of a system with the operator $T \in \mathcal{T}$, then $r(\tau) = T[s(\tau)]$. Now suppose that $\bar{s}(\tau)$ is an approximation to $s(\tau)$, then the signal $\bar{r}(\tau) = T^1[\bar{s}(\tau)]$ will be an approximation to $r(\tau)$.

It will now be shown that an operator from $\mathcal{T}^1$ is a gain. Let $\bar{s}(\tau) = \bar{S}_1\varphi_1(\tau)$ and $\bar{r}(\tau) = \bar{R}_1\varphi_1(\tau)$, then $\bar{R}_1\varphi_1(\tau) = T_P T[\bar{S}_1\varphi_1(\tau)]$. Since $\bar{R}_1$ is the projection of $T[\bar{S}_1\varphi_1(\tau)]$ onto the basis function $\varphi_1(\tau)$, (1) yields

$$\bar{R}_1 = \int_{-\infty}^{t} T[\bar{S}_1 \ \varphi_1(\tau)]\varphi_1(\tau)d\tau$$

Defining G as

$$G = \frac{1}{\bar{S}_1} \int_{-\infty}^{t} T[\bar{S}_1\varphi_1(\tau)] \ \varphi_1(\tau)d\tau \qquad (3)$$

yields $\bar{R}_1 = G \ \bar{S}_1$ and $\bar{R}_1 \ \varphi_1(\tau) = G \ \bar{S}_1\varphi_1(\tau)$. That is, the operator $T^1 = T_P T$ is the gain G defined by (4). Equation (4) makes it evident that G is, in general, a function of $\bar{S}_1$; G will be independent of $\bar{S}_1$ when T is a linear operator.

The gain G is the exponential describing function corresponding to the operator T. The name is derived from the choice of $\varphi_1(\tau)$. Taking $\{\varphi_1(\tau)\}$ to be the reversed time Laguerre functions on the interval $(-\infty,t]$, $\varphi_1(\tau)$ is the growing exponential $\sqrt{2\gamma} \ e^{\gamma(\tau-t)}$.

One of the reasons the exponential describing function appears to be very promising as an analysis tool is the ease with which G can be determined in many cases. As an example, consider the linear system with the operator T_H defined by

$$r(\tau) = \int_{-\infty}^{\tau} h(\tau-\nu) \ s(\nu)d\nu$$

where $h(\nu)$, equal to zero for $\nu < 0$, is absolutely integrable on the interval $[0,+\infty)$. Let $H(\lambda)$ be the Laplace transform of $h(\nu)$, then (3) yields

$$G_H = H(\gamma) \qquad (4)$$

As another example, consider the memoryless, nonlinear system with the operator T_N defined by

$$r(\tau) = N[s(\tau)]$$

where $N(s)$ is a function satisfying a Lipschitz condition for all s. With the observation $\varphi_1(\tau)d\tau = (1/\gamma)d\varphi_1(\tau)$, equation (3) yields

$$G_N = \frac{1}{\gamma\bar{S}_1^{\ 2}} \int_{0}^{\sqrt{2\gamma} \ \bar{S}_1} N(\sigma)d\sigma \qquad (5)$$

FEEDBACK SYSTEMS ANALYSIS

The analysis of a feedback system is a major goal in the development of an analysis tool. In this section, therefore, the exponential describing function will be demonstrated as a tool in the analysis of a class of feedback systems. Let T_H and T_N be the operators introduced in the last section, then the block diagram in Figure 1 depicts the class of feedback systems to be examined.

In the application of an analysis technique which will yield an approximate solution to a problem, there must exist prior knowledge that the problem actually has a solution; otherwise, the existence of an approximate solution has no meaning. Thus, the first item to be studied is that of existence of a unique solution to the problem. The technique to be employed has been used by other authors [1,2,5-11]; it is the signal space which is different in this paper.

Solution Existence

Let $\mathcal{L}_2(-\infty,t]$ be a Banach space and let $\|q\|$ denote the norm of $q \in \mathcal{L}_2(-\infty,t]$. The usual notation $\|T\|$ will be used to denote the norm of a linear operator $T \in \mathcal{T}$.

A mapping of $\mathcal{L}_2(-\infty,t]$ into itself by an operator $T \in \mathcal{T}$ is said to be a contraction, if there exists a number $k < 1$ that

$$\|Tq_1 - Tq_2\| < k\,\|q_1 - q_2\| \tag{6}$$

for any two elements q_1, $q_2 \in \mathcal{L}_2(-\infty,t]$. If it is a contraction-mapping, the contraction-mapping fixed-point theorem [3] establishes the fact that there is a unique element $q \in \mathcal{L}_2(-\infty,t]$ such that $q = Tq$.

The feedback system under investigation is described by the equation

$$r = T_H s - T_H T_N r \tag{7}$$

To establish reasonable conditions for the existence of a unique system response, r, it is expedient to put the feedback system equation in a different form. Add $\mu T_H r$ to both sides of (7); assuming that the linear operator $[I + \mu T_H]$ possesses an inverse on $\mathcal{L}_2(-\infty,t]$, the feedback system equation becomes

$$r = [I + \mu T_H]^{-1} T_H s - [I + \mu T_H]^{-1} T_H [T_N - \mu I]r \tag{8}$$

Equation (8) is of the form $r = Mr$, where $M \in \mathcal{T}$. Observe that

$$\|Mr_1 - Mr_2\| \leqslant \|[I + \mu T_H]^{-1} T_H\|\, \|[T_N - \mu I]r_1 - [T_N - \mu I]r_2\|$$

Let the Lipschitz condition satisfied by N(s) be

$$\alpha(s_1 - s_2) \leqslant N(s_1) - N(s_2) \leqslant \beta\,(s_1 - s_2)$$

where $s_1 \geqslant s_2$ and the constants α and β are such that $\alpha + \beta \geqslant 0$. Then

$$\|[T_N - \mu I]r_1 - [T_N - \mu I]r_2\| \leqslant \eta(\mu)\|r_1 - r_2\|$$

where $\eta(\mu) = \max \left\{|\beta-\mu|, |\mu-\alpha|\right\}$ and

$$\|Mr_1 - Mr_2\| \leq \eta(\mu)\| [I + \mu T_H]^{-1} T_H\| \|r_1 - r_2\|$$

Let $\mu_o = (\alpha+\beta)/2$. Sandberg [12] has shown that

$$\eta(\mu_o) \|[I + \mu_o T_H]^{-1} T_H\| \leq \eta(\mu)\| [I + \mu T_H]^{-1} T_H\|$$

if $\eta(\mu)\|[I - \mu T_H]^{-1} T_H\| < 1$. Since $\eta(\mu_o) = (\beta-\alpha)/2$, a sufficient condition for M to be a contraction-mapping is

$$\frac{\beta-\alpha}{2} \|[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H\| < 1 \qquad (9)$$

Approximate Solution Existence

Assume that the sufficient conditions for the existence of a unique response, r, are satisfied. Then, it becomes a meaningful task to establish the existence of an approximate solution. This will be accomplished without imposing additional conditions.

To obtain the feedback system defined on $\mathcal{L}^1_2(-\infty,t]$ which corresponds to the original feedback system of Figure 1, each operator is replaced by its corresponding operator from $\mathcal{T}^1$ and the stimulus signal is replaced by its approximation $T_P s \in \mathcal{L}^1_2(-\infty,t]$. Thus, the approximate response $\bar{r}$, must satisfy the equation

$$\bar{r} = T_P T_H T_P s - T_P T_H T_P T_N \bar{r} \qquad (10)$$

Observe that $T_P \bar{r} = \bar{r}$ and that $T_P T_H T_P = T_H T_P$. Note: In [1,2], the papers providing the initial guidelines for this paper, the projection operator commutes with the linear operator. In this paper, this is not true; that is $T_H T_P \neq T_P T_H$. With these identities, the addition of $\mu_o T_H \bar{r}$ to both sides of (10) will yield

$$[I + \mu_o T_H]\bar{r} = T_P T_H T_P s - T_P T_H T_P [T_N - \mu_o I]\bar{r}$$

By assumption $[I + \mu_o T_H]$ possesses an inverse on $\mathcal{L}_2(-\infty,t]$. Now, $[I + \mu_o T_H]^{-1} \in \mathcal{T}$ is a linear operator such that $[I + \mu_o T_H]^{-1} = T_P[I + \mu_o T_H]^{-1}$ on $\mathcal{L}^1_2(-\infty,t]$. Therefore $[I + \mu_o T_H]^{-1} \in \mathcal{T}^1$, and

$$\bar{r} = T_P [I + \mu_o T_H]^{-1} T_H T_P s - T_P [I + \mu_o T_H]^{-1} T_H T_P [T_N - \mu_o I]\bar{r} \qquad (11)$$

Equation (11) is of the form $\bar{r} = M^1\bar{r}$, where $M^1 \in \mathcal{T}^1$.

The operator M^1 is also a member of $\mathcal{T}$ by definition of $\mathcal{T}^1$. Since $\|T_P\| = 1$, it is a simple matter to show that (9) is a sufficient condition for M^1 to be a contraction-mapping on $\mathcal{L}_2(-\infty,t]$. The fixed point of the mapping is in $\mathcal{L}^1_2(-\infty,t]$; hence, (11) possesses a unique solution.

Present Time Approximate Response

Suppose that $\bar{r}$ exists, then (10) has a solution. In particular, it has a solution at the present time t. This is the solution value of interest.

Set $G_H = T_P T_H$ and $G_N = T_P T_N$ and let $\bar{s} = T_P s$. To avoid subsequent notation difficulties, set $\mathscr{s}(t) = \bar{s}(t)$ and $\mathscr{r}(t) = \bar{r}(t)$, then (10) yields

$$[\mathscr{s} - G_H^{-1}\mathscr{r}] = G_N \mathscr{r} \qquad (12)$$

The gain G_H is a constant equal to $H(\gamma)$ and the gain G_N is, in general, a function of $\mathscr{r}$. For each value of $\mathscr{s}$, (12) is an algebraic or transcendental equation for $\mathscr{r}$ expressed as the equality of two functions of $\mathscr{r}$. Equation (12) may be solved, therefore, by ordinary graphical methods.

The value of $\mathscr{s}(t)$ is equal to $S_1(t)\varphi_1(t) = \sqrt{2\gamma}\ S_1(t)$. With $S_1(t)$ obtained by (1), it is clear that

$$\mathscr{s}(t) = 2\gamma \int_{-\infty}^{t} e^{\gamma(\nu - t)}\ s(\nu)d\nu \qquad (13)$$

Together, (12) and (13) yield the approximation, $\mathscr{r}(t)$, to the actual response, $r(t)$.

Set $t = \tau$ in (13), then $\mathscr{s}(\tau) \in \mathscr{L}_2(-\infty, t]$ is an approximation to $s(\tau)$ obtained by the mapping $T_{\mathscr{P}} \in \mathscr{T}$ and defined by (13). Equation (12) or the equation

$$\mathscr{r} = [1 + \frac{\alpha+\beta}{2} G_H]^{-1} G_H \mathscr{s} - [1 + \frac{\alpha+\beta}{2} G_H]^{-1} G_H [G_N - \frac{\alpha+\beta}{2}]\mathscr{r} \qquad (14)$$

as derived from (11), will yield the corresponding $\mathscr{r}(\tau)$. This observation will be put to use shortly.

Selection of γ

The value of γ, the neper frequency of the growing exponential $\varphi_1(\tau)$, has not yet been specified. This will be done in conjunction with an examination of the approximation error.

Let $T \in \mathscr{T}$ be a contraction-mapping on the Banach space $\mathscr{L}_2(-\infty, t]$. Let q_o be arbitrary element of $\mathscr{L}_2(-\infty, t]$ and let $q_m = T^m q_o$. In the proof of the contraction-mapping fixed-point theorem the inequality

$$\|q - q_m\| \leqslant \frac{k^m}{1-k}\| q_1 - q_o\| \qquad (15)$$

where $q = Tq$, is shown to be valid.

The contraction-mapping under investigation is defined by (8) with

$$k = \frac{\beta-\alpha}{2}\| [I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H\|$$

Set $m = 0$ and take $r_o(\tau)$ to be $\mathscr{r}(\tau)$; then, by (15) with (8) and (14)

$$\|r - \mathscr{r}\| \leqslant \frac{1}{1-k} \left\{ \|[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H - [1 + \frac{\alpha+\beta}{2} G_H]^{-1} G_H T_{\mathscr{P}}\| \; \|s\| \right.$$
$$+ \| \{[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H [T_N - \frac{\alpha+\beta}{2} I]$$

$$-[1 + \frac{\alpha+\beta}{2} G_H]^{-1} G_H [G_N - \frac{\alpha+\beta}{2}]\} \mathscr{r} \|\Big\} \qquad (16)$$

Observe that

$$\|[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H - [1 + \frac{\alpha+\beta}{2} G_H]^{-1} G_H T_\rho\| \qquad (17)$$

is a function of a γ since $G_H = H(\gamma)$. The neper frequency γ will be selected so to minimize the quantity (17).

Selecting γ in this way will not, in general, minimize this bound on the approximation error $\|r - \mathscr{r}\|$, since the other terms in (16) are also functions of γ, unless they are identically zero. An attempt at selecting γ such as to minimize this bound on $\|r - \mathscr{r}\|$ would, most likely, be frustrated by the enormity of the task. The minimization of (17) represents a rationalized compromise. The rationalization is: The operator $[T_N - \frac{\alpha+\beta}{2} I]$ has a maximum slope of $(\beta-\alpha)/2$ and a minimum slope of $-(\beta-\alpha)/2$. Suppose the average of the maximum and minimum slopes -- zero -- is indicative of the gain of a nominally equivalent linear operator with constant gain. Then, it would be reasonable to suppose the error bound contribution from the term

$$[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H [T_N - \frac{\alpha+\beta}{2} I] \mathscr{r}$$

will be small. Since $[G_N - \frac{\alpha+\beta}{2}]$ is the exponential describing function for $[T_N - \frac{\alpha+\beta}{2} I]$, it would appear reasonable to also suppose that the error bound contribution from

$$[1 + \frac{\alpha+\beta}{2} G_H]^{-1} G_H [G_N - \frac{\alpha+\beta}{2}] \mathscr{r}$$

will be small. Thus, it is not unreasonable to suppose that the term (17) is dominant in this error bound and that minimization of (17) will reduce this error bound.

Example I

The first feedback system to be subjected to analysis, to illustrate the exponential describing function as an analysis tool, is shown in Figure 2. In this example, the space $\mathcal{L}_2(-\infty,t]$ will be $L_2(-\infty,t]$.

For the nonlinearity $N(r) = \tanh r$, the values of α and β are 0 and 1 respectively; therefore, $\mu_o = (\alpha+\beta)/2 = 1/2$. The Laplace transform

$$1 + \frac{\alpha+\beta}{2} H(\lambda) = 1 + \frac{1}{2} \; \frac{10^5}{\lambda+10^5}$$

is positive for $\mathcal{R}e\,[\lambda] \geqslant 0$. Thus, the operator $[I + \frac{\alpha+\beta}{2} T_H]$ possesses a bounded, causal inverse on $L_2(-\infty,t]$. This fact may be proven in a manner similar to that used by Sandberg, in the first part of his proof of lemma 5 in [13], with the added information from Theorem XVIII in [14].

The norm of the operator $[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H$ is

$$\|[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H\| = \sup_{\omega} |[1 + \frac{1}{2} H(j\omega)]^{-1} H(j\omega)| = \frac{2}{3}$$

This fact is derived through use of Parseval's theorem concerning Fourier transforms; the technique of proof may be seen in [7,9]. It is obvious that the inequality (9) is satisfied and, hence, that the feedback system of Figure 2 possesses a unique response $r(\tau)\in L_2(-\infty,t]$ for all $s(\tau)\in L_2(-\infty,t]$.

The quantity (17) is

$$\sup_{\omega} | [1 + \frac{1}{2} H(j\omega)]^{-1} H(j\omega) - [1 + \frac{1}{2} H(\gamma)]^{-1} H(\gamma) \frac{2\gamma}{j\omega+\gamma} |$$

Upon setting $\gamma = (3/2) \times 10^5$, this quantity is easily shown to be identically zero, and, hence, to be a minimum.

The gains G_H and G_N are

$$G_H = \frac{2}{5}$$

$$G_N = 2 \frac{\text{Ln} \cosh \mathscr{r}}{\mathscr{r}^2}$$

The gain G_N follows from (5) and $\mathscr{r} = \sqrt{2\gamma}\ \bar{R}_1$. For this example, (12) is

$$[\mathscr{s} - \frac{5}{2}\mathscr{r}] = 2 \frac{\text{Ln} \cosh \mathscr{r}}{\mathscr{r}} \qquad (17)$$

a graphical solution of (17) yields the $\mathscr{r}$ versus $\mathscr{s}$ relationship depicted in Figure 3. The assymptotes for large and small $\mathscr{s}$ are shown as dashed lines.

A sine wave commencing at time zero will be the input signal; that is

$$\begin{aligned} s(\tau) &= 0 & \tau &\leqslant 0 \\ &= \sigma \sin \omega\tau & 0 &\leqslant \tau \leqslant t \end{aligned}$$

for the particular values $\sigma = 10$ and $\omega = (3/2) \times 10^5$, the approximation $\mathscr{s}(t)$ given by (13) is

$$\begin{aligned} \mathscr{s}(t) &= 0 & t &\leqslant 0 \\ &= \sqrt{2} \times 10 \{ \sin [(3/2) \times 10^5 t - 0.786] + \exp[(3/2) \text{x} 10^5 t] \} & 0 &\leqslant t \end{aligned}$$

The $\mathscr{r}$ versus $\mathscr{s}$ curve of Figure 3 together with $\mathscr{s}(t)$ yields $\mathscr{r}(t)$. A suitable approximation to the feedback system output of this particular example, $r'(t)$, would be $\tanh \mathscr{r}(t)$. The approximate response, $\tanh \mathscr{r}(t)$, is shown in Figure 4; the actual output, $r'(t)$, as obtained by machine calculation, is also shown.

An additional observation is: The response $r(t)$ is ultimately periodic; this may be verified by application of a theorem given by Sandberg and Beneš in [15]. The approximation $\mathscr{r}(t)$ correctly indicates this result; it is also ultimately periodic

Example II

The second illustrative example will produce an approximate analysis of the feedback system depicted in Figure 1 with

$$H(\lambda) = 10^2/\lambda$$

and

$$N(r) = r \qquad |r| \leqslant 1$$

$$= 9r - 8 \qquad |r| \geqslant 1$$

The space $\mathcal{L}_2(-\infty,t]$ will be the space of measurable functions, $q(\tau)$, bounded as

$$|q(\tau)| \leqslant K\, e^{\delta\tau} \qquad \text{almost everywhere}$$

with $K < \infty$ and δ a positive constant less than γ. With the norm

$$\|q\| = \operatorname*{ess\,sup}_{\tau\in(-\infty,t]} |q(\tau)\, e^{-\delta\tau}|$$

$\mathcal{L}_2(-\infty,t]$ is a Banach space. To show that $\mathcal{L}_2(-\infty,t]$ is a Banach space, it is necessary to show that every Cauchy sequence in $\mathcal{L}_2(-\infty,t]$ converges to an element of $\mathcal{L}_2(-\infty,t]$; this fact is a direct consequence of the completeness of $L_\infty(-\infty,t]$ [4]. Note: In this example $h(\nu)$ is not absolutely integrable; it is sufficient for $T_H \in \mathcal{T}$ that $h(\nu)e^{-\delta\nu}$ be absolutely integrable on the interval $[0,+\infty)$, which it is.

The nonlinearity $N(r)$ establishes the values of 1 and 9 for α and β, respectively; thus, $\mu_o = (\alpha+\beta)/2 = 5$. The Laplace transform

$$1 + \frac{\alpha+\beta}{2} H(\lambda) = 1 + 5\,\frac{10^2}{\lambda}, \quad \mathrm{Re}[\lambda] \geqslant \delta$$

is positive for $\mathrm{Re}[\lambda] \geqslant \delta$. Therefore, the operator $[I + \frac{\alpha+\beta}{2} T_H]$ has a bounded, causal inverse on $\mathcal{L}_2(-\infty,t]$. The proof of this fact follows the proof given by Sandberg for theorem I(a) of [7] after observing that $q(\tau)e^{-\delta\tau} \in L_\infty(-\infty,t]$ for $q(\tau) \in \mathcal{L}_2(-\infty,t]$ and noting Theorem XVIII of [14].

Let $\mathcal{L}^{-1}\{[1 + \frac{\alpha+\beta}{2} H(\lambda)]^{-1} H(\lambda)\}$ denote the inverse Laplace transform of $[1 + \frac{\alpha+\beta}{2} H(\lambda)]^{-1} H(\lambda)$, $\mathrm{Re}[\lambda] \geqslant \delta$. Then the norm of the operator $[I + \frac{\alpha+\beta}{2} T_H]^{-1} T_H$ is

$$\int_0^\infty | \mathcal{L}^{-1}\{[1 + 5H(\lambda)]^{-1} H(\lambda)\}|\, e^{-\delta\nu}\, d\nu = \frac{10^2}{5\mathrm{x}10^2+\delta}$$

This fact is established as the immediate consequence of the definition for the norm of a linear operator. Inequality (9) is clearly satisfied and, thus the feedback system possesses a unique response $r(\tau) \in \mathcal{L}_2(-\infty,t]$ for all $s(\tau) \in \mathcal{L}_2(-\infty,t]$.

It is a simple matter to show that the quantity (17) is minimized — in fact identically zero — for $\gamma = 5 \times 10^2$. The gains G_H and G_N are

$$G_H = \frac{1}{5}$$

and

$$G_N = 1 \qquad |\rho| \leqslant 1$$

$$= \frac{9\rho^2 - 16\rho + 8}{\rho^2} \qquad |\rho| \geqslant 1$$

Equation (12) is easily solved; the resulting ρ versus ς relationship is shown in Figure 5.

Let the feedback system stimulus signal, $s(\tau)$, be the pulse

$$s(\tau) = 0 \qquad \tau \leqslant \min\,[0,t]$$

$$= \sigma \qquad 0 < \tau \leqslant \min\,[1,t]$$

$$= 0 \qquad 1 < \tau \leqslant t$$

The approximation $\varsigma(t)$ is

$$\varsigma(t) = 0 \qquad t \leqslant 0$$

$$= 2\sigma[1-e^{-\gamma t}] \qquad 0 < t \leqslant 1$$

$$= 2\sigma[e^{\gamma}-1]e^{-\gamma t} \qquad 1 < t$$

Figure 5 together with $\varsigma(t)$ yield $\rho(t)$; $\rho(t)$ and $r(t)$, as obtained by machine calculation, are shown in Figure 6 for σ equal to 1,3,and 10.

The growing exponential bound on the response signal with respect to the time variable t raises the question: Ultimately is the feedback system stable, in some sense of the word? It has been shown [16] that $r(t)$ will be square integrable on the interval $[0,+\infty)$.

Other Feedback Systems

The primary goal of this section has been to validate the use of the exponential describing function in the analysis of a class of feedback systems — those of Figure 1. These are not the only systems for which the analysis can be justified with ease. For example, the class of systems created by interchanging the linear and nonlinear operators may be subjected to analysis under precisely the same set of validating conditions; this fact follows from a study similar to that in this section. The intent of these comments is: The exponential describing function should not be viewed as an analysis tool with applicability only to systems of the type shown in Figure 1.

CONCLUDING DISCUSSION

The stated objectives for this paper have been met. The exponential describing function has been developed. Conditions under which a system, from a particular class of systems, has a unique solution have been established as a justification for submitting the system to an approximate analysis using the exponential describing function. The examples give the reader an opportunity to evaluate the application of the developed tool.

The rationalization by which minimization of (17), with respect to γ, as a basis for selecting γ is intuitively based. Future results, by the author and/or the readers, may provide a basis for selection of γ established upon fewer subjective evaluations.

Suppose that minimization of the quantity (17) is meaningful. Then, the implication is that the operator $[I + \frac{\alpha+\beta}{2} T_H]^{-1}T_H$ should have the qualitative characteristics of a low-pass filter, since the operator T_γ, corresponds to a low-pass filter.

The exponential describing function as an analysis tool was the result of signal approximation by projection on the sub-space $\mathcal{L}_2^1(-\infty,t] \subset \mathcal{L}_2(-\infty,t]$. The reader wishing to pursue the consequences -- to analysis -- of signal approximation by projection on $\mathcal{L}_2^n(-\infty,t]$, n not necessarily unity, is referred to the solution approximations stemming from Galerkin's method [17]. More accurate approximations are possible for n sufficiently large, but the effort needed to achieve that approximation will be significantly greater than that required when n = 1.

Since the goal of this paper was to develop the exponential describing function and justify its use in an abstract Banach space contained in $L_2(-\infty,t]$, existence of the inverse of $[I + \mu T_H]$ on $\mathcal{L}_2(-\infty,t]$ was by assumption. In each of the examples, the conditions insuring the existence of an inverse were given and comments on how proofs might be effected were made. This appeared to be advisable since conditions for the existence of the inverse can be established by techniques similar to those used by other authors for different spaces. In addition to the two papers by Sandberg [7,13] already cited, the reader is referred to [1,2,11,13,15,18,19].

The ultimate behavior of the feedback system response, mentioned at the end of each example, is quite important and has been under extensive investigation in recent years. In addition to the two papers [15,16] mentioned in conjunction with the examples, the reader is referred to [11,13,18,20-23] for some interesting results.

REFERENCES

[1] Sandberg, I.W., On the Response of Nonlinear Control Systems to Periodic Input Signals, Bell System Technical Journal, Vol. XLIII, No. 3, May 1964 pp. 911-926,

[2] Sandberg, I.W. On Truncation Techniques in the Approximate Analysis of Periodically Time-Varying Nonlinear Networks, IEEE Transactions on Circuit Theory, Vol. CT-11, No. 2, June 1964 pp. 195-201.

[3] Kolmogorov, A.N. and Fomin, S.V., Elements of the Theory of Functions and Functional Analysis, Rochester, New York: Graylock Press, 1957.

[4] Taylor, Angus E., Introduction to Functional Analysis, New York: John Wiley and Sons, 1963.

[5] Sandberg, I.W., On the Properties of Some Systems that Distort Signals-I, Bell System Technical Journal, Vol. XLII, No. 5, September 1963, pp. 2033-2046.

[6] Sandberg, I.W., On the Properties of Some Systems that Distort Signals-II, Bell System Technical Journal, Vol. XLIII, No. 1, Part 1, January 1964, pp. 91-112.

[7] Sandberg, I.W., Signal Distortion in Nonlinear Feedback Systems, Bell System Technical Journal, Vol. XLII, No. 6, November 1963, pp. 2533-2550.

[8] Desoer, Charles A., Nonlinear Distortion in Feedback Amplifiers, IEEE Transactions on Circuit Theory, Vol. CT-9, No. 1, March 1962, pp. 2-6.

[9] Zames, George, Nonlinear Operators for System Analysis, Technical Report 370, Research Laboratory of Electronics, MIT, Cambridge, Massachusetts, August 25, 1960.

[10] Zames, George, Functional Analysis Applied to Nonlinear Feedback Systems, IEEE Transactions on Circuit Theory, Vol. CT-10, No. 3, September 1963, pp. 392-404.

[11] Zames, George, On the Input-Output Stability of Nonlinear, Time-Varying Feedback Systems, Preprint of Paper Delivered at the National Electronics Conference, Chicago, Illinois, October 1964.

[12] Sandberg, I.W., A Note on the Application of the Contraction-Mapping Fixed-Point Theorem to a Class of Nonlinear Functional Equations, SIAM Review, Vol. 7, No. 2, April 1965, pp. 199-204.

[13] Sandberg, I.W., On the $\mathcal{L}_2$-Boundedness of Solutions of Nonlinear Functional Equations, Bell System Technical Journal, Vol. XLIII, No. 4, Part 2, July 1964, pp. 1581-1600.

[14] Paley, Raymond E.A.C. and Wiener, Norbert, Fourier Transforms in the Complex Domain, New York: American Mathematical Society, 1934.

[15] Sandberg, I.W. and Beneš, V.E., On the Properties of Nonlinear Integral Equations that Arise in the Theory of Dynamical Systems, Bell System Technical Journal, Vol. XLIII, No. 6, November 1964, pp. 2839-2853.

[16] Sandberg, I.W., Some Results on the Theory of Physical Systems Governed by Nonlinear Functional Equations, Bell System Technical Journal, Vol. XLIV, No. 5, May-June 1965, pp. 871-898.

[17] Krasnosel'skii, M.A., Topological Methods in the Theory of Nonlinear Integral Equations, New York: The MacMillan Company, 1964, pp. 169-180.

[18] Zames, George, Conditions for the Input-Output Stability of Nonlinear, Time-Varying Systems, Derived Using the Concept of Loop Gain, Preprint available from author at M.I.T., Cambridge, Massachusetts.

[19] Leon, B.J. and Anderson, D.R., The Stability of Pumped Nonlinear Reactance Circuits, IEEE Transactions on Circuit Theory, Vol. CT-10, No.4, December 1963, pp. 468-476, Corrected in Vol. CT-11, No. 1, March 1965, pp. 143-144.

[20] Sandberg, I.W., A Frequency-Domain Condition for the Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element, Bell System Technical Journal, Vol. XLIII, No. 4, Part 2, July 1964, pp. 1601-1608.

[21] Sandberg, I.W., A Condition for the $\mathcal{L}_\infty$-Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element, Bell System Technical Journal, Vol. XLIII, No. 4, Part 2, July 1964, pp. 1815-1817.

[22] Sandberg, I.W., On the Boundedness of Solutions of Nonlinear Integral Equations, Bell System Technical Journal, Vol. XLIV, No. 3, March 1965, pp. 439-454.

[23] Sandberg, I.W., An Observation Concerning the Application of the Contraction-Mapping Fixed-Point Theorem, and a Generalization of a Result Concerning the Norm-Boundedness of Solutions of Nonlinear Functional Equations, Bell System Technical Journal, Vol.XLIV, October 1965.

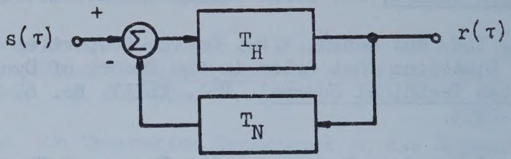
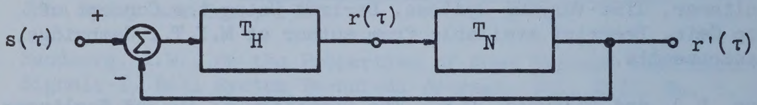


Figure 1 Feedback System


+
s(τ)
-
Σ
T_H
r(τ)
T_N
r'(τ)

$$H(\lambda) = \frac{10^5}{\lambda+10^5}$$

$$N(r) = \tanh r$$

Figure 2 Low-Pass Amplifier with Saturating Output Stage

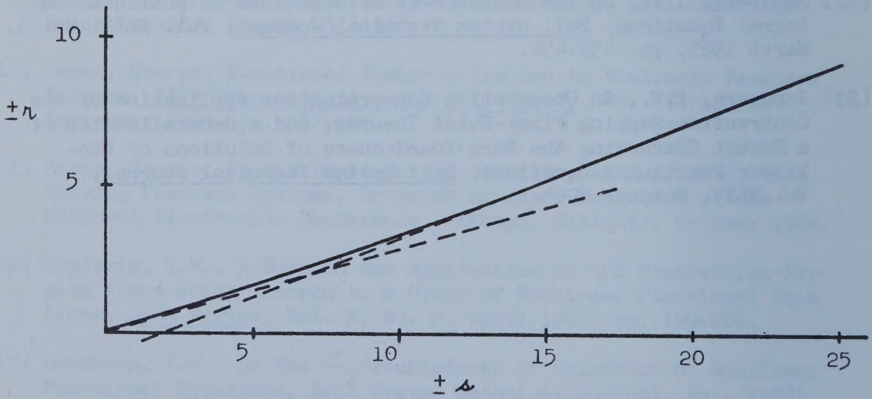


Figure 3 Example I —— r versus s

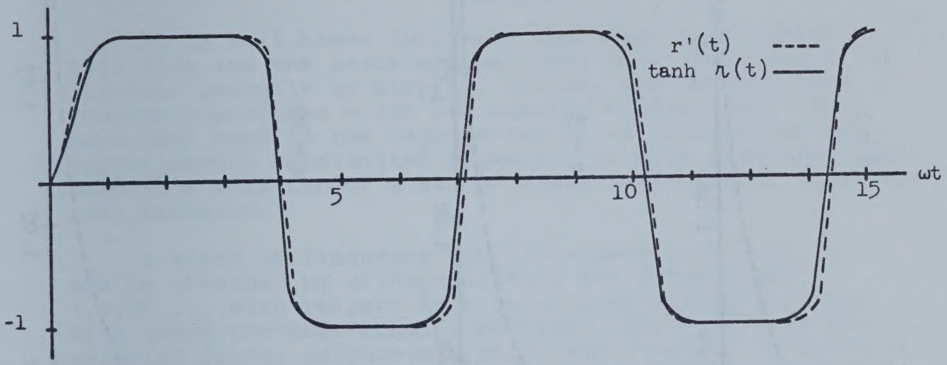


Figure 4 Example I —— $r'(t)$ and tann $r(t)$

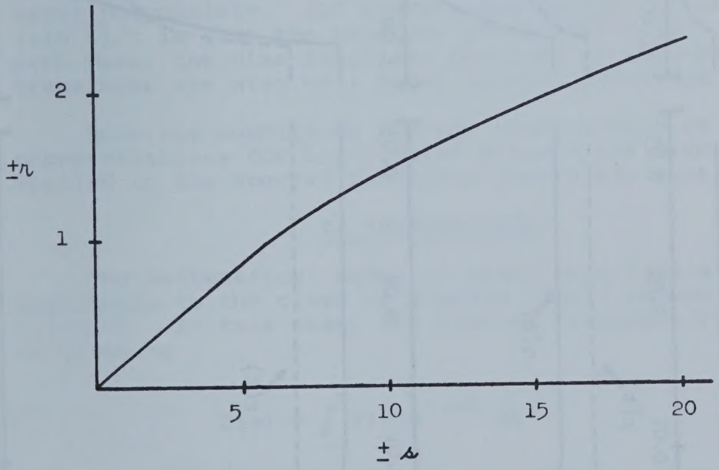


Figure 5 Example II —— r versus s

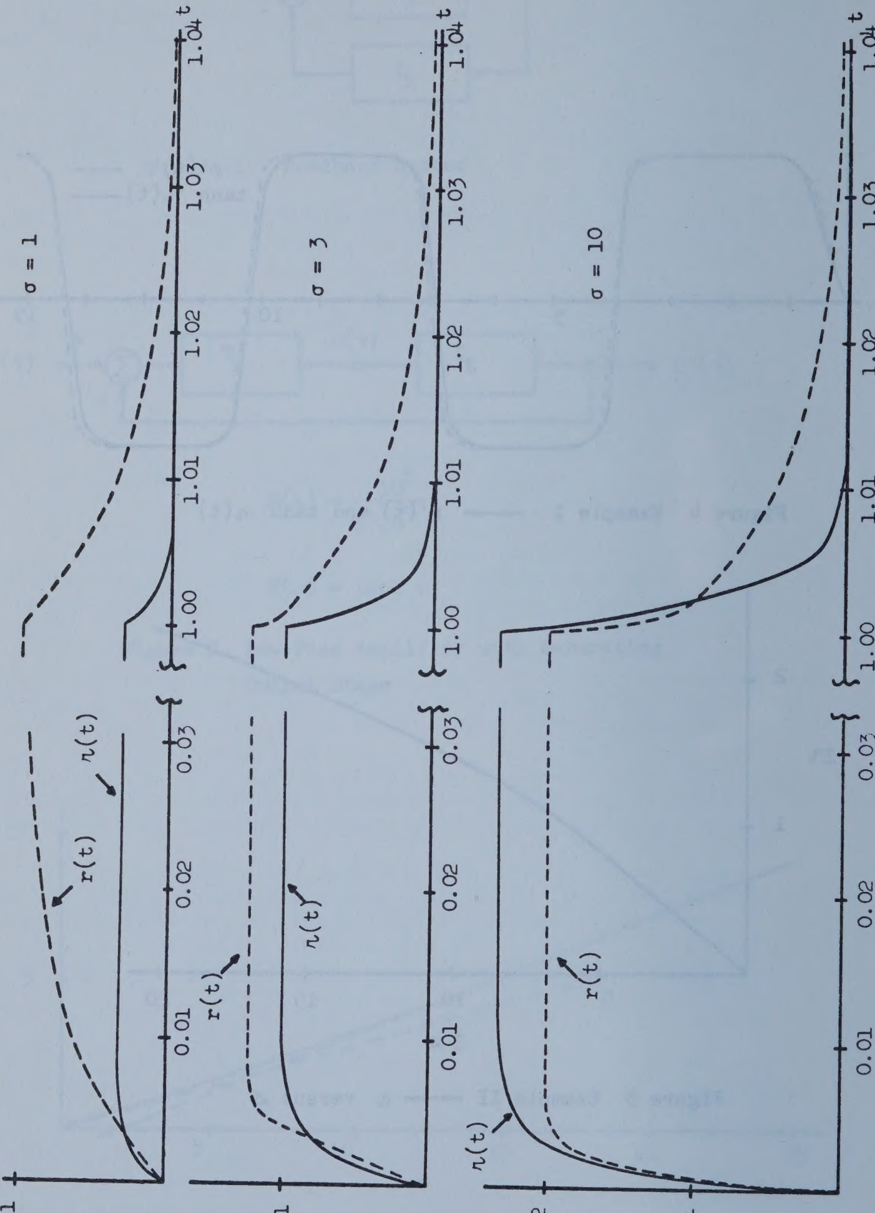


Figure 6 Example II —— r(t) and 𝓇(t) for σ Equal to 1,3 and 10

ON BANDLIMITED PROPERTIES OF FOURIER TRANSFORM PAIRS OF SOME SPECIAL FUNCTIONS

K. YAO and J. B. THOMAS
Department of Electrical Engineering
Princeton University
Princeton, New Jersey

ABSTRACT

It is well known that both the sampling function (sin t)/t and the prolate spheriodal wave functions $\Psi_i(t)$ [treated recently by Slepian, Pollak, and Landau] have Fourier transforms which are exactly bandlimited. Both have been used in the representation of bandlimited and approximately bandlimited signals. In each case the transform is a solution of a second order self-adjoint differential equation.

A class of frequency functions bandlimited to (-1,1) can be obtained by orthogonalizing the polynomials $1, \omega, \omega^2, \ldots$ with respect to a certain weighting function $W(\omega)$ using the Gram-Schmidt process. The result is the weighted Jacobi polynomials which are complete in $L_2(-1,1)$ and are also the solutions of a second order self-adjoint differential equation. Special cases of the Jacobi polynomials include the Gegenbauer polynomials, the Legendre polynomials, the Tschebyscheff polynomials, and the Tesseral polynomials. [Of course, the Fourier transform of (sin t)/t is just the Legendre polynomial $P_o(\omega)$.] In each case, the time functions obtained as inverse Fourier transforms are also well known special functions.

Relevant conditions for the construction of sampling representations for bandlimited signals are discussed and applied to the special functions previously mentioned.

I. INTRODUCTION

The mathematical model of exact band-limitation is applicable to the class of signals $\{f(t)\}$ belonging to $L_2(-\infty,\infty)$. In this case, the Fourier transform $F(\omega)$ of f(t), as given by

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt \quad , \qquad (1)$$

exists and is assumed to be zero outside of (-1,1) radians/sec., i.e.

$$F(\omega) = 0 \quad , \qquad \omega \notin (-1,1) \quad . \qquad (2)$$

It is frequently desirable to express the signals satisfying

Eqs. (1) and (2) in terms of other sets of functions in both the time and frequency domains. Thus any signal satisfying Eqs. (1) and (2) can be written as

$$F(\omega) = \sum_{n=-\infty}^{\infty} a_n \psi_n(\omega) \quad , \quad \omega \in (-1,1) \tag{3}$$

where $\{\psi_n(\omega)\}$ is any complete orthonormal set in $L_2(-1,1)$ so that

$$\int_{-1}^{1} \psi_n(\omega)\overline{\psi_m(\omega)}d\omega = \delta_{mn} \tag{4}$$

and

$$a_n = \int_{-1}^{1} F(\omega)\overline{\psi_n(\omega)}d\omega \tag{5}$$

The right side of Eq. (3) converges in the mean to the frequency function $F(\omega)$.

If $\phi_n(t)$ is the inverse Fourier transform of $\psi_n(\omega)$, given by

$$\phi_n(t) = \frac{1}{2\pi}\int_{-1}^{1} \psi_n(\omega)e^{i\omega t}d\omega \quad , \tag{6}$$

then the set of functions $\{\phi_n(t)\}$ is complete and orthonormal in $L_2(-\infty,\infty)$ [1]. A mean-convergent representation of f(t) is given by

$$f(t) = \sum_{n=-\infty}^{\infty} a_n \phi_n(t) \quad , \quad t \in (-\infty,\infty) \tag{7}$$

where the coefficient a_n as obtained by

$$a_n = \int_{-\infty}^{\infty} f(t)\overline{\phi_n(t)}\, dt \tag{8}$$

is equal to the a_n given by Eq. (5) through the Plancherel Theorem [2].

In general, additional constraints or specifications of the coefficients or the functions $\psi_n(\omega)$ are necessary if additional properties of the expansion are desired. The set of coefficients, $\{a_n\}$, and the set of complete orthonormal functions, $\{\psi_n(\omega)\}$, are related for each n by Eq. (5). If the set $\{a_n\}$ is required to have the property that a_n is proportional to $f(t_n)$, then the set $\{\psi_n(\omega)\}$ is specified through the relationship

$$cf(t_n) = a_n = \int_{-1}^{1} F(\omega)\overline{\psi_n(\omega)}d\omega \tag{9}$$

This requirement that the coefficient of expansion, a_n, be proportional to the value of $f(t)$ at time t_n is a meaningful constraint since these sample value coefficients are easy to obtain from the actual signal. The resulting representation in the time domain

$$f(t) = \sum_{n=-\infty}^{\infty} cf(t_n)\varphi_n(t) \quad , \quad t \in (-\infty,\infty) \tag{10}$$

will be called a sampling expansion or sampling theorem.

II. A CLASS OF FOURIER TRANSFORM PAIRS OF BAND-LIMITED SIGNALS

Let us consider the Legendre equation given by [3]

$$\frac{d}{d\omega}\left[(1-\omega^2)\frac{d\psi(\omega)}{d\omega}\right]+n(n+1)\psi(\omega) = 0 \tag{11}$$

This is a second order self-adjoint differential equation having solutions for integer n which are continuous (in fact, polynomials of order n) in the interval $|\omega| \leq 1$. The solutions, for integer n, are the Legendre polynomials

$$\psi_n(\omega) = P_n(\omega) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n, \; n = 0,1,2,..., \tag{12}$$

which are complete and orthogonal in (-1,1). In particular, for n = 0,

$$P_o(\omega) = 1 \qquad |\omega| \leq 1 \tag{13}$$

The inverse Fourier transform of this frequency function is

$$\mathfrak{F}^{-1}\{P_o(\omega)\} = \frac{1}{2\pi}\int_{-1}^{1} e^{j\omega t}d\omega = \frac{1}{\pi}\frac{\sin t}{t} \tag{14}$$

It is well known that this function is related to the ordinary Sampling Theorem or Shannon Series [4,5] for a function f(t) band-limited in ω to (-1,1) by

$$f(t) = \sum_{k=-\infty}^{\infty} f(k\pi) \frac{\sin(t-k\pi)}{t-k\pi} \qquad (15)$$

As a second example, consider the second order self-adjoint differential equation [6]

$$\frac{d}{d\omega}\left[(1-\omega^2)\frac{d\psi(\omega)}{d\omega}\right] + (\chi-\omega^2)\psi(\omega) = 0 \qquad (16)$$

obtained by adding the term $-\omega^2\psi(\omega)$ to the Legendre equation given by Eq. (11). For the set of eigenvalues, $\{\chi_n\}$, $n = 0,1,2,\ldots$, where $0<\chi_o<\chi_1<\ldots$, there corresponds a set of functions

$$\psi_n(\omega) = S_{on}(\omega) \;, \quad n = 0,1,2,\ldots, \qquad (17)$$

complete and orthogonal in $L_2(-1,1)$ called the angular prolate spheriodal functions of the first kind. The inverse Fourier transforms of the $\psi_n(\omega) = S_{on}(\omega)$ are called the radial prolate spheriodal functions and differ from the angular functions only by a real scale factor so that $\phi_n(t) = A_n S_{on}(t)$, where A_n is a constant. The two sets, $\{S_{on}(\omega)\}$ and $\{S_{on}(t)\}$, $n = 0,1,2,\ldots$, are simultaneously complete and orthogonal in $L_2(-1,1)$ and $L_2(-\infty,\infty)$, respectively. The basic properties of the prolate spheriodal functions are given in [6], [7], and [8]. Many of the optimum energy concentration properties of these functions useful in signal theory are given in [9] and [10]. Specifically, the normalized function $S_{oo}(\omega)$ has the largest energy in the time interval $(-1,1)$ of all unit energy functions band-limited to $(-1,1)$ in the ω-domain. It is interesting to note, however, that the coefficient of expansion, a_n, given by Eq. (5) is not easily calculated.

We now consider the second order differential equation

$$\frac{d}{d\omega}\left[(1-\omega^2)\frac{dy(\omega)}{d\omega}\right]+\left[\beta-\alpha-(\alpha+\beta)\omega\right]\frac{dy(\omega)}{d\omega} +n(n+\alpha+\beta+1)y = 0 \qquad (18)$$

For $\alpha = \beta = 0$, it is apparent that this equation reduces to the Legendre equation given by [11]. The solutions for integer n are well known [11] and are called the Jacobi polynomials $P_n^{(\alpha,\beta)}(\omega)$ defined by

$$y(\omega) = P_n^{(\alpha,\beta)}(\omega) = \frac{(-1)^n(1-\omega)^{-\alpha}(1+\omega)^{-\beta}}{2^n n!}\frac{d^n}{d\omega^n}\left[(1-\omega)^{n+\alpha}(1+\omega)^{n+\beta}\right] \qquad (19)$$

for $\alpha>-1$, $\beta>-1$, $n = 0,1,2,\ldots$. These polynomials are orthogonal on the interval $(-1,1)$ with respect to the weight function

$$W(\omega) = (1-\omega)^{\alpha}(1+\omega)^{\beta} \quad , \quad \alpha>-1,\ \beta>-1 \tag{20}$$

so that

$$\int_{-1}^{1} P_n^{(\alpha,\beta)}(\omega) P_m^{(\alpha,\beta)}(\omega)(1-\omega)^{\alpha}(1+\omega)^{\beta} d\omega = C_n^2 \delta_{mn} \tag{21}$$

The Jacobi polynomials are the functions that result from orthogonalizing the polynomials $1, \omega, \omega^2, \ldots$ on the interval $(-1,1)$ with respect to $W(\omega)$ by means of the Gram-Schmidt process. Although Eq. (18) is not self-adjoint, the associated self-adjoint equation is obtained by multiplying Eq. (18) by the weight function $W(\omega)$ and is given by

$$\frac{d}{d\omega}\left[(1-\omega^2)(1-\omega)^{\alpha}(1+\omega)^{\beta} \frac{dy(\omega)}{d\omega}\right]$$

$$+ n(n+\alpha+\beta+1)(1-\omega)^{\alpha}(1+\omega)^{\beta} y(\omega) = 0 \tag{22}$$

Particular cases of the Jacobi polynomials have been studied extensively and have special names. The Gegenbauer polynomial, denoted by $T_n^{\beta}(\omega)$, is the special case of the Jacobi polynomial $P_n^{(\alpha,\beta)}(\omega)$ when $\alpha = \beta$. The following names are given to some special classes of Gegenbauer polynomials:

$\beta = 0$: Legendre polynomials, $P_n(\omega)$

$\beta = -\frac{1}{2}$: Tschebyscheff polynomials of the first kind, $T_n^{(-1/2)}(\omega)$

$\beta = \frac{1}{2}$: Tschebyscheff polynomials of the second kind, $T_n^{(1/2)}(\omega)$

$\beta = m$ (pos. integer): Tesseral polynomials, $T_n^m(\omega)$.

In order to represent the band-limited signal $F(\omega)$, let

$$\psi_n(\omega) = P_n^{(\alpha,\beta)}(\omega)(1-\omega)^{\alpha}(1+\omega)^{\beta}/C_n$$

Then

$$F(\omega) = \sum_{n=0}^{\infty} \frac{a_n}{C_n} P_n^{(\alpha,\beta)}(\omega)(1-\omega)^{\alpha}(1+\omega)^{\beta} \tag{23}$$

where

$$a_n = \frac{1}{C_n} \int_{-1}^{1} F(\omega) P_n^{(\alpha,\beta)}(\omega)\, d\omega \tag{24}$$

The inverse Fourier transform of $\psi_n(\omega)$ is given by

$$\phi_n(t) = \frac{1}{2\pi C_n} \int_{-1}^{1} P_n^{(\alpha,\beta)}(\omega)(1-\omega)^{\alpha}(1+\omega)^{\beta} e^{i\omega t} d\omega \tag{25}$$

This last expression can be written with the aid of Eq. (19) as

$$\phi_n(t) = A_{1n} \int_{-1}^{1} e^{i\omega t} \frac{d^n}{d\omega^n} \left[(1-\omega)^{n+\alpha}(1+\omega)^{n+\beta} \right] d\omega \tag{26}$$

where A_{1n} is a constant. After integration by parts n times, Eq. (26) becomes

$$\phi_n(t) = A_{2n} t^n \int_{-1}^{1} e^{i\omega t}(1-\omega)^{n+\alpha}(1+\omega)^{n+\beta} d\omega \tag{27}$$

The confluent hypergeometric function ${}_1F_1(a;c;z)$ is given as [12]

$${}_1F_1(a;c;z) = \frac{\Gamma(c) 2^{1-c} e^{z/2}}{\Gamma(a)\Gamma(c-a)} \int_{-1}^{1} e^{zu/2}(1-u)^{c-a-1}(1+u)^{a-1} du \tag{28}$$

Let $c = 2n+\alpha+\beta+2$, $a = n+\beta+1$, $u = \omega$, and $z = 2it$. Then Eq. (27) reduces to

$$\phi_n(t) = A_{3n} t^n e^{-it} {}_1F_1(n+\beta+1;\ 2n+\alpha+\beta+2;\ 2it) \tag{29}$$

It is well known [13] that the Whittaker's form of solution of the confluent hypergeometric differential equation is the Whittaker function $M_{\varkappa,\mu}(z)$ given by

$$M_{\varkappa,\mu}(z) = e^{-z/2} z^{c/2} {}_1F_1(a;c;z) \tag{30}$$

where $\varkappa = -a + c/2$ and $\mu = (c-1)/2$. Then

$$M_{\frac{\alpha-\beta}{2},\,\frac{2n+\alpha+\beta+1}{2}}(t) = e^{-it}(2it)^{\frac{2n+\alpha+\beta+2}{2}}\;{}_1F_1(n+\beta+1;2n+\alpha+\beta+2;2it) \tag{31}$$

and

$$\phi_n(t) = A_{4n}M_{\frac{\alpha-\beta}{2},\,\frac{2n+\alpha+\beta+1}{2}}(t) \; . \tag{32}$$

Thus the inverse Fourier transform of the weighted Jacobi polynomial is proportional to Whittaker's function. For $\alpha = \beta$, the inverse Fourier transform of the weighted Gegenbauer polynomial is given by

$$\phi_n(t) = A_{5n}t^n e^{-it}\;{}_1F_1(n+\beta+1;\; 2(n+\beta+1);\; 2it) \tag{33}$$

For $\varkappa = 0$, $\mu = \nu$ and $z = 2it$, the Whittaker function $M_{\varkappa,u}(z)$ reduces to the Bessel function of the first kind, $J_\nu(t)$. Since

$$J_\nu(t) = \frac{t^\nu e^{-it}}{2^\nu\Gamma(\nu+1)}\;{}_1F_1(\nu+\frac{1}{2}\;;\; 2(\nu+\frac{1}{2});\; 2it) \tag{34}$$

Eq. (33) reduces to

$$\phi_n(t) = \frac{A_{6n}J_{n+\beta+(1/2)}(t)}{t^{\beta+(1/2)}} \tag{35}$$

In the case where $\beta = 0$, the inverse Fourier transform of the weighted Legendre polynomial is given by

$$\phi_n(t) = \frac{A_{7n}J_{n+(1/2)}(t)}{t^{1/2}} \tag{36}$$

Since the spherical Bessel function $j_n(t)$ is related to the Bessel function of the first kind by

$$j_n(t) = (\frac{\pi}{2})^{(1/2)}\frac{J_{n+(1/2)}(t)}{t^{1/2}} \tag{37}$$

then Eq. (36) reduces to

$$\phi_n(t) = A_{7n}j_n(t) \tag{38}$$

The various transform relationships discussed in this section are summarized in Table I on the next page.

The band-limited signal f(t) is given by Eq. (7) and its transform F(ω) by Eq. (23). The coefficients, a_n, are given by Eq. (24) as well as by

$$a_n = \frac{1}{A_{8n}} \int_{-\infty}^{\infty} f(t) t^n e^{-it} {}_1F_1(n+\beta+1; 2n+\alpha+\beta+2; 2it)\, dt \tag{39}$$

These coefficients are not simply related to F(ω) or f(t). In practice, it is very difficult to evaluate either Eqs. (24) or (39). We will now discuss some relevant conditions for the construction of sampling representation for band-limited signals.

III. SAMPLING EXPANSIONS FOR BAND-LIMITED SIGNALS

Let $\{\psi_n(\omega)\}$ and $\{\phi_n(t)\}$ be complete and orthonormal in $L_2(-1,1)$ and $L_2(-\infty,\infty)$, respectively. If the coefficient, a_n is required to be $\sqrt{2\pi}$ times the sample value of f(t) at time t_n, then, from Eq. (5),

$$f(t_n) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} F(\omega) \overline{\psi_n(\omega)} d\omega \tag{40}$$

Since the inverse Fourier transform of F(ω) is f(t), it is clear that the set of complete orthonormal functions

$$\psi_n(\omega) = (2)^{-(1/2)} e^{-i\pi\omega n} \quad , \; n = \ldots, -1, 0, 1, \ldots, \tag{41}$$

is sufficient for the sampling expansion of signals band-limited to (-1,1). The result of using this given set in Eqs. (3) and (6) yields [14] the Shannon Sampling Theorem of Eq. (15).

Now, instead of requiring that all the members of $\{\psi_n(\omega)\}$ be orthonormal in $L_2(-1,1)$, let us require a second set $\{e^{it_n\omega}\}$, to be orthonormal to the set $\{\psi_n(\omega)\}$. That is

$$\int_{-1}^{1} e^{it_n\omega} \psi_m(\omega) d\omega = \delta_{mn} \tag{42}$$

Then the coefficient a_n is given by

$$a_n = \int_{-1}^{1} F(\omega) e^{it_n\omega} d\omega = 2\pi f(t_n) \tag{43}$$

The properties of such biorthogonal sets are studied in [15] and [16]. It can be shown [17] that the corresponding sampling expansion of signals band-limited to (-1,1) can be given by

TABLE I - FOURIER TRANSFORM PAIRS

	parameter	Time Domain - $\phi_n(t) \in L_2(-\infty,\infty)$	Frequency Domain - $\psi_n(\omega) \in L_2(-1,1)$
		$C\phi_n(t) = \frac{1}{2\pi}\int_{-1}^{1} \psi_n(\omega) e^{i\omega t} d\omega$	$\psi_n(\omega) = \int_{-\infty}^{\infty} C\phi_n(t) e^{-i\omega t}\, dt$
1.	c	Order zero radial prolate spheroidal function of the first kind - $R_{on}^{(1)}(c,t)$	Order zero angular prolate spheroidal function of the first kind - $S_{on}(c,\omega)$
2.	α,β	Whittaker function of the first kind - $M_{\frac{\alpha-\beta}{2},\frac{2n+\alpha+\beta+1}{2}}(t)$	Weighted Jacobi polynomial of the first kind - $(1-\omega)^{\alpha}(1+\omega)^{\beta} P_n^{(\alpha,\beta)}(\omega)$
3.	$\alpha=\beta$	Weighted Bessel function of the first kind, order $(n+\beta+1/2)$- $t^{-\beta-1/2} J_{n+\beta+1/2}(t)$	Weighted Gegenbauer polynomial of the first kind - $(1-\omega^2)^{\beta} P_n^{\beta}(\omega)$
4.	$\alpha=\beta=0$	Spherical Bessel function of the first kind - $j_n(t)$	Legendre polynomial of the first kind - $P_n(\omega)$
5.	$\alpha=\beta=-\frac{1}{2}$	Bessel function of the first kind, order n - $J_n(t)$	Weighted Tschebyscheff polynomial of the first kind - $(1-\omega^2)^{-1/2} P_n^{(-1/2)}(\omega)$
6.	$\alpha=\beta=\frac{1}{2}$	Weighted Bessel function of the first kind, order n+1 - $t^{-1} J_{n+1}(t)$	Weighted Tschebyscheff polynomial of the second kind - $(1-\omega^2)^{1/2} P_n^{(1/2)}(\omega)$
7.	$\alpha=\beta=m$	Weighted Bessel function of the first kind, order n+m+1/2- $t^{-m-1/2} J_{n+m+1/2}(t)$	Weighted Tesseral polynomial of the first kind - $(1-\omega^2)^{m} P_n^{m}(\omega)$

$$f(t) = \sum_{n=-\infty}^{\infty} f(t_n) \frac{G(t)}{G'(t_n)(t-t_n)} \tag{44}$$

where $|t_n - n\pi| < D = (\ln 2)/\pi$, all integer n, and

$$G(t) = \prod_{n=-\infty}^{\infty} (t-t_n) \tag{45}$$

It is known from Kramer [18] and Campbell[19] that if G(x) is a member of the class of functions which vanishes outside of (a,b) and is in $L_2(a,b)$, then f(t) is representable in a sampling representation

$$f(t) = \sum_{n=-\infty}^{\infty} f(t_n)\phi(t,t_n) \tag{46}$$

Here f(t) is related to G(x) by the integral transform

$$f(t) = \int_a^b G(x)\psi(x,t)\,dx \tag{47}$$

where the kernel of the transform has the property that $\{\psi(x,t_n)\}$, $n = \ldots,-1,0,1,\ldots$, is complete and orthonormal in $L_2(a,b)$. The sampling expansion in Eq. (46) follows from the expansion

$$G(x) = \sum_{n=-\infty}^{\infty} a_n \overline{\psi(x,t_n)} \tag{48}$$

where $a_n = f(t_n)$ and the function $\phi(t,t_n)$ is given by

$$\phi(t,t_n) = \int_a^b \overline{\psi(x,t_n)}\,\psi(x,t)\,dx \tag{49}$$

In this formulation, although f(t) is related to G(x) by Eq. (47), it is not clear how to obtain G(x) from f(t). Bochner [20] gives the necessary and sufficient conditions for the representation of a pair of unitary transformations in the Hilbert space $L_2(-\infty,\infty)$ by a pair of integral transformations with kernels $\psi(x,t)$ and $\overline{\psi(t,x)}$. Some examples of this procedure for deriving sampling theorems have been given [17,18,19].

ACKNOWLEDGMENT

The authors wish to express their appreciation to the National Science Foundation and to the U.S. Army Research Office - Durham for their generous support of this research under Grant GP-1647 and Contract DA-31-124-ARO-D-292, respectively.

REFERENCES

1. R.E.A.C. Paley, N. Wiener, Fourier Transforms in the Complex Domain, American Mathematical Society, Providence, R.I., 1943, p. 29.
2. R.E.A.C. Paley, N. Wiener, ibid. p. 2.
3. N.N. Lebedev, Special Functions and Their Applications, (Translated from the Russian by R.A. Silverman), Prentice-Hall, Inc., Englewood Cliffs, N.J., 1965, Chaps. 4 and 7.
4. E.T. Whittaker, "On the Functions Which Are Represented by the Expansions of the Interpolation Theory", Proc. Royal Soc., Edinburgh, Vol. 35, 1915, pp. 181-194.
5. C.E. Shannon, "Communication in the Presence of Noise", Proc. IRE, Vol. 37, 1949.
6. P.M. Morse, H. Feshbach, Methods of Theoretical Physics, McGraw-Hill Book Co., Inc., New York, 1953, Chaps. 5, 10, 11.
7. C. Flammer, Spheroidal Wave Functions, Stanford University Press, Stanford California, 1957.
8. J.A. Stratton, P.M. Morse, L.J. Chu, J.D.C. Little, and F.J. Corbato, Spheroidal Wave Functions, The Technology Press of M.I.T., Cambridge Mass., and John Wiley and Sons, New York, 1956.
9. D. Slepian, H.O. Pollak, "Prolate Spheriodal Wave Functions, Fourier Analysis and Uncertainty - I", Bell Sys. Tech. J., Vol. 40, January 1961, pp. 43-64.
10. H.J. Landau, H.O. Pollak, "Prolate Spheriodal Wave Functions, Fourier Analysis and Uncertainty - II", Bell Sys. Tech. J., Vol. 40, January 1961, pp. 65-84.
11. G. Szego, Orthogonal Polynomials, American Mathematical Society Publication Vol. XXIII, American Mathematical Society, New York, 1959, p. 67.
12. W. Magnus, F.Oberhettinger, Formulas and Theorems for the Functions of Mathematical Physics, Chelsea Pub., New York, 1954, pp. 86-88.
13. A. Erdelyi, W. Magnus, F. Oberhettinger, F.G. Tricomi, Higher Trancendental Functions, Vol. I, McGraw-Hill Book Co., Inc., New York, N.Y., 1953, p. 264.
14. C. Shannon, "A Mathematical Theory of Communication", Bell Sys. Tech. J., Vol. 27, 1948, pp. 379-423, 623-659.
15. N. Levinson, Gap and Density Theorems, American Mathematical Society Colloquiem Pub., New York, 1940.
16. F. Riesz, B. Sz-Nagy, Functional Analysis, Frederick Ungar Pub. Co., New York, 1955, Section 86.
17. K. Yao, On Some Representations and Sampling Expansions for Band-Limited Signals, Doctoral Dissertation, Princeton University, Princeton, N.J., April 1965.
18. H.P. Kramer, "A Generalized Sampling Theorem", Journal of Mathematics and Physics, Vol. 38, 1959, pp. 68-72.
19. L.J. Campbell, "A Comparison of the Sampling Theorems of Kramer and Whittaker", J. Soc. Indus. App. Math., Vol. 12, No. 1, 1964, pp. 117-130.
20. S. Bochner, "Inversion Formulae and Unitary Transformations", Annals of Math., Vol. 35, 1934, pp. 111-115.

FOURIER TECHNIQUES FOR THE RESOLUTION OF OVERLAPPING WAVEFORMS

J. R. REEDER* AND S. S. WOLFF
THE JOHNS HOPKINS UNIVERSITY, Baltimore, Maryland

ABSTRACT

Many physical processes produce signals composed of overlapping waveforms or echoes. Such signals arise in the areas of seismology, nuclear magnetic resonance, electron paramagnetic resonance, infrared and microwave spectroscopy, electrophoresis, and multipath communications. Efforts to decompose these signals into their component waveforms or wavelets date from the early work of Lord Rayleigh in 1871 [1] to the recent publications of Medgyessy [2] and Tukey [3].

In this paper we demonstrate the underlying Fourier nature of many solutions to the decomposition problem. The effectiveness of the decomposition procedures in the presence of additive noise is considered. In particular, the results of Root [5] on the singular detection problem of statistical communication theory demonstrate the possibility of an unambiguous resolution of a noisy signal by the fundamental method of Medgyessy.

The method of "quefrency alanysis" which has been proposed by Tukey [3] is presented and the relationship to the fundamental method of Medgyessy is discussed. The possibility of an unambiguous resolution of a noisy signal by this method is also demonstrated.

The decomposition procedures are illustrated by the resolution of a simple signal taken from the field of spectroscopy.

II. INTRODUCTION AND FORMULATION OF THE PROBLEM

The resolution problem concerns signals that are superpositions of basic waveforms whose structural form is approximately known. A mathematical model appropriate to many applications is of the form

$$v(t) = s(t) + n(t) \tag{2-1}$$

$$s(t) = \sum_{i=1}^{N} A_i \phi(\mu_i, \lambda_i, t) \tag{2-2}$$

where $v(t)$ is the observed signal, $s(t)$ is an ideal signal predicted by an underlying physical theory, and $n(t)$ is the departure from the

---

This work was supported in part by The Office of Naval Research under contract Nonr-4010 (13) awarded to The Johns Hopkins University.

* Mr. Reeder is a doctoral candidate on the Westinghouse Aerospace Electrical Division Doctoral Program.

ideal signal caused by random measurement errors, unavoidable additive noise, or inadequacies of the physical theory.

The ideal signal $s(t)$ is a superposition of N basic waveforms (or wavelets) $\phi(\mu_i, \lambda_i, t)$. The time of occurrence (or epoch) for each wavelet is defined by the set of parameters μ_i and the structure relative to the epoch is determined by the set of parameters λ_i. We do not treat the problem in its full generality here, but consider instead an important special case commonly occurring in the fields of seismology, spectroscopy, and electrophoretical research. In this case, λ_i is a single parameter regulating the time scale of the i th wavelet and the epoch parameter μ_i is a simple translation parameter. The simplified form for $s(t)$ is then

$$s(t) = \sum_{i=1}^{N} A_i \Phi\left(\frac{t - \mu_i}{\lambda_i}\right) \qquad (2\text{-}3)$$

Teicher [12] and Robbins [13] have demonstrated that a unique decomposition of $s(t)$ is not possible for arbitrary types of wavelets. In the following sections we assume the form of the wavelets is such that a unique decomposition of $s(t)$ is possible, although ambiguity may exist in decomposition of the experimental signal $v(t)$ due to the presence of the additive noise $n(t)$.

III. RESOLUTION BY FOURIER TECHNIQUES

The Fourier transforms of two signals which are related by a pure translation differ only by an exponential factor proportional to the translation between the signals according to the relation

$$\Phi(\omega; \mu) = \mathcal{F}[\phi(t - \mu)] = e^{-j\omega\mu} \mathcal{F}[\phi(t)] = e^{-j\omega\mu} \Phi(\omega) \qquad (3\text{-}1)$$

where $\mathcal{F}$ represents the Fourier integral operator. For the simplified case where all the wavelets comprising the ideal signal $s(t)$ given by equation (2-3) possess identical structural parameters λ, the Fourier transform $S(\omega; \lambda)$ takes the simple form

$$S(\omega; \lambda) = \mathcal{F}\left[\sum_{i=1}^{N} A_i \Phi\left(\frac{t - \mu_i}{\lambda}\right)\right] = \Phi(\omega; \lambda) \sum_{i=1}^{N} A_o e^{-j\omega\mu_i} \qquad (3\text{-}2)$$

This transform consists of one factor $\Phi(\omega; \lambda)$ which is a function solely of the structural form of a primitive wavelet $\phi\left(\frac{t}{\lambda}\right)$ and a second factor

$$B(\omega) = \sum_{i=1}^{N} A_i e^{-j\omega\mu_i} \qquad (3\text{-}3)$$

containing information concerning the weights and epochs of the component wavelets. The factor $B(\omega)$ may be uniquely recovered from equation (3-2) providing $\Phi\ (\omega;\ \lambda)$ is known a priori and does not vanish:

$$B(\omega) = \Phi^{-1}(\omega;\ \lambda) \cdot S(\omega;\ \lambda) \tag{3-4}$$

In the event that $\Phi\ (\omega;\ \lambda)$ vanishes on some ω -set, Bracewell [9] demonstrates that equation (3-4) determines $B(\omega)$ uniquely only where $\Phi(\omega;\ \lambda)$ is supported. Many practical wavelets such as the Gaussian and Lorentzian wavelets in spectroscopy, possess transforms $\Phi(\omega;\ \lambda)$ which are everywhere non-zero. For simplicy of presentation we henceforth assume all wavelets considered here to possess transforms of this type.

When the number N of wavelets comprising $s(t)$ is known, an estimate of the epoch and weight parameters may be obtained by approximating the expression (3-4) for $B(\omega)$ by the form given in equation (3-3). The method of least-squares minimization or an adaptation of the Prony method of exponential approximation proposed by Stone [8] is appropriate for the estimation of the parameters.

The inverse Fourier transform $b(t)$ of equation (3-3) and hence of equation (3-4) for the case when $\Phi\ (\omega;\ \lambda)$ is precisely known consists of N weighted impulses of the form

$$b(t) = \mathcal{F}^{-1}\left[\Phi^{-1}(\omega;\ \lambda)S(\omega;\ \lambda)\right] = \sum_{i=1}^{N} A_i\ \delta(t - \mu_i) \tag{3-4A}$$

This method of decomposition is equivalent to the fundamental method of Pal Medgyessy [2] for the separation of mixtures of probability density functions. The method has the advantage that the number N of wavelets comprising $s(t)$ need not be known a priori. It has the further advantage that under certain conditions (Ref. 2, pp. 58-59) the structural parameter of the wavelets, if unknown, may be determined by rewriting equation (3-4) in terms of a parameter ρ yielding

$$b(t;\ \rho) = \mathcal{F}^{-1}\left[\Phi^{-1}(\omega;\ \rho) \cdot S(\omega;\ \lambda)\right] \tag{3-5}$$

In this iterative scheme, ρ is gradually increased from zero until $b(t;\ \rho)$ approaches the string of weighted impulses of equation (3-4) as $\rho \uparrow \lambda$. Clearly this same technique may be employed when the wavelets comprising $s(t)$ are characterized by several different values of $\lambda, \lambda_1 < \lambda_2 < \cdots \lambda_m$: as the wavelets characterized by λ_1 are resolved, they are subtracted from $s(t)$. The parameter ρ is further increased to search for λ_2, and so on.

The decomposition procedure indicated in equation (3-4) may also be accomplished directly by time domain operations on $s(t)$. We use the fact that multiplication of $S(\omega)$ by any function of ω corresponds uniquely to some linear operation upon $s(t)$; for example:

$$\omega^n S(\omega) \longleftrightarrow (-j)\ \frac{d^n s(t)}{dt^n} \tag{3-6}$$

and

$$e^{-j\omega\Delta} S(\omega) \longleftrightarrow s(t - \Delta) \qquad (3\text{-}7)$$

Substitution of the approximation

$$\Phi^{-1}(\omega;\ \lambda) = \sum_{n=0}^{\infty} c_n \Psi_n(\omega) \qquad (3\text{-}8)$$

into equation (3-4) allows the inverse transform $b(t)$ corresponding to $B(\omega)$ to be expressed directly in terms of a set of operators $\theta^{(n)}$ acting upon $s(t)$

$$b(t) = \sum_{n=0}^{\infty} c_n \theta^{(n)} s(t) \qquad (3\text{-}9)$$

The functions $\{\Psi_n\}$ are chosen to correspond to operations which are easily performed upon $s(t)$. Different choices for $\{\Psi_n\}$ lead to superficially different (though fundamentally equivalent) methods of decomposition. Substantive differences of the methods are caused, however, by the practical necessity of truncating the series (3- 8) after a finite number of terms. A number of examples are presented below illustrating some of the common methods for approximating $\Phi^{-1}(\omega;\ \lambda)$

Example 1:

The differentiation method of Eddington [10] is extensively used in spectroscopy for resolution enhancement of experimental spectra and for correction of distortions introduced into the spectra by the measuring apparatus. This method results from choosing

$$\Psi_n(\omega) = \omega^n \qquad (3\text{-}10)$$

which yields

$$b(t) = c_o s(t) + \sum_{n=1}^{\infty} (-j)^n c_n \frac{d^n s(t)}{dt^n} \qquad (3\text{-}11)$$

Example 2:

The differentiations specified by equation (3-11) must normally be replaced by their finite difference approximations for computations involving experimental signals. Noting that

$$\sin \frac{\omega\Delta}{2} S(\omega) = \frac{s(t + \Delta/2) - s(t - \Delta/2)}{2j} = \frac{1}{2j} D_\Delta^{(1)} s(t) \qquad (3\text{-}12)$$

where $D_\Delta^{(1)}$ is the first finite difference operator, leads to the method of finite differences due to Lord Rayleigh [1] wherein

$$\Psi_n(\omega) = \left(\sin\frac{\omega\Delta}{2}\right)^n \qquad (3\text{-}13)$$

so that

$$b(t) = c_o s(t) + \sum_{n=1}^{\infty} \frac{c_n}{(2j)^n} D_{\Delta}^{(n)} s(t) \qquad (3\text{-}14)$$

Example 3:

Closely related to the methods of Eddington and Lord Rayleigh is the method of displacements obtained by choosing

$$\Psi_n(\omega) = e^{-jn\omega\Delta} \qquad (3\text{-}15)$$

so that

$$b(t) = c_o s(t) + \sum_{n=1}^{\infty} c_n s(t - n\Delta) \qquad (3\text{-}16)$$

An appropriate manner for choosing the constants $\{c_n\}$ in the above examples is determination by either a Taylor expansion or by the least squares approximation of $\Phi^{-1}(\omega; \lambda)$ using a small number of terms of the series (3-11) and evaluating the constants $\{c_n\}$ by least-squares minimization.

Example 4:

The methods of decomposition presented in the above examples have been intimately related to the derivative operator. In contrast, Berger and van Cittert [11] have proposed a decomposition procedure whose basic operation is one of integration based upon the method of successive substitutions from the theory of integral equations. In developing this method it is convenient to consider the signal $s(t)$ given by equation (2-3) for the case of identical structural parameters λ as the convolution of a basic waveform $\phi(t; \lambda)$ with a train of weighted impulses $b(t)$ given by equation (3-4A)

$$s(t) = \int_{-\infty}^{\infty} b(\tau) \cdot \phi(t - \tau; \lambda) d\tau = b(t) * \phi(t; \lambda) \qquad (3\text{-}17)$$

The decomposition of $s(t)$ is thus equivalent to solving this integral equation for $b(t)$. We use the method of successive substitutions for the solution, beginning with $s(t)$ as the initial approximation to $b(t)$

$$b_o(t) = s(t) \qquad (3\text{-}18)$$

The method proceeds by choosing the discrepancy between $s(t)$ and the convolution of $\phi(t; \lambda)$ with $b\ (t)$ as the first correction to $b(t)$

$$b_1(t) = b_o(t) + [s(t) - \phi(t; \lambda)] * s(t) \qquad (3\text{-}19)$$

The procedure is then iterated to obtain an nth approximation to $b(t)$

$$b_n(t) = b_{n-1}(t) + [s(t) - \phi(t;\lambda)] * b_{n-1}(t) \qquad (3\text{-}20)$$

which may also be written as

$$b_n(t) = s(t) + \sum_{i=1}^{N} e_i \qquad (3\text{-}21)$$

where

$$e_1 = s(t) - \phi(t;\lambda) * s(t) = [\delta(t) - \phi(t;\lambda)] * s(t)$$

$$e_i = [\delta(t) - \phi(t;\lambda)] * e_{i-1}$$

The Fourier transform of equation (3-21) yields

$$B(\omega) = \left[1 + [1 - \Phi(\omega;\lambda)] + [1 - \Phi(\omega;\lambda)]^2 + \ldots [1 - \Phi(\omega;\lambda)]^n\right] S(\omega)$$

or, in the limit (3-22)

$$B(\omega) = \sum_{n=0}^{\infty} [1 - \Phi(\omega;\lambda)]^n \, S(\omega) = \frac{S(\omega)}{1 - [1 - \Phi(\omega;\lambda)]}$$

This method of decomposition is thus equivalent to expanding $\Phi^{-1}(\omega;\lambda)$ of equation (3-4) in a geometrical series. A sufficient condition for the convergence of this expansion is $|1 - \Phi(\omega;\lambda)| < 1$ for all ω.

IV. AN ADAPTATION OF TUKEY'S METHOD

In the preceeding section we noted that the Fourier transform of the ideal signal $s(t)$ may be filtered to yield a transform $B(\omega)$ (3-3) consisting of a superposition of periodic functions in the frequency variable ω. A decomposition of $s(t)$ was then afforded by performing a harmonic analysis of $B(\omega)$ in which each harmonic (or quefrency) corresponds to the epoch of a wavelet. Tukey [3] has proposed the somewhat analogous procedure of "quefrency alanysis" for the resolution of overlapping stochastic waveforms utilizing the power spectral density of the signal. We present the method here as adapted for pulse signals.

The method proceeds by estimating the power spectral density $P_{ss}(\omega)$ of the signal $s(t)$

$$P_{ss}(\omega) = \frac{1}{T} U_{\phi\phi}(\omega) \cdot U_{bb}(\omega) \qquad (4\text{-}1)$$

where T is the duration of the signal $s(t)$ and $U_{\phi\phi}(\omega)$ and $U_{bb}(\omega)$ are the energy densities of $\phi(t)$ and $b(t)$ respectively. Taking the logarithm of the power spectral density yields

$$\ln P_{ss}(\omega) = \ln \frac{U_{\phi\phi}(\omega)}{T} + \ln U_{bb}(\omega) \qquad (4\text{-}2)$$

Since $B(\omega)$ is a periodic function of ω as given by (3-3), $\ln U_{bb}(\omega) = \ln|B(\omega)|^2$ is also a periodic function of ω whose quefrency composition is determined by the epoch structure of the wavelets comprising $s(t)$. Then $\ln P_{ss}(\omega)$ is the sum of a periodic function containing the information concerning the epochs and weights of the wavelets and an aperiodic function $U_{\phi\phi}(\omega)$ which depends only upon the basic pulse shape $\phi(\frac{t}{\lambda})$. Subtraction of the term $\ln U_{\phi\phi}(\omega)/T$ (analogous to multiplication by $\Phi^{-1}(\omega; \lambda)$ in the Medgyessy method) leaves the periodic function $U_{bb}(\omega)$. The inverse Fourier transform of $\ln U_{bb}(\omega)$ is a string of weighted impulses containing the resolution information concerning the wavelets. Unlike the Medgyessy technique, however, the harmonic analysis of $\ln U_{bb}(\omega)$ is complicated by extraneous quefrencies introduced by the quadratic nature of $U_{bb}(\omega)$ and the operation of extracting its logarithm. The analysis may be simplified at this point by taking the antilogarithm of $\ln U_{bb}(\omega)$. Thus while the resolution information is somewhat confounded, it is nevertheless recoverable.

The essential attribute of the Tukey technique is the logarithmic operation which provides an additive decoupling of the basic pulse shape from the information concerning the weights and epochs of the wavelets. Thus even when the pulse shape is imperfecty known, or completely unknown as in the case of a stochastic signal, an estimate of $\ln U_{bb}(\omega)$ is still obtainable as the periodic portion of $\ln P_{ss}(\omega)$.

V. THE NOISY CASE

When the resolution techniques treated earlier are applied to noisy data, it is natural to expect that the ultimate resolution attainable will be limited by the random fluctuations. We now show that there is an analogue of the singular case in statistical detection theory that allows perfect resolution even when noise is present.

For simplicity, and without loss of generality, we assume that only two wavelets are to be resolved; thus

$$v(t) = a_1 \phi(t - \mu_1) + a_2 \phi(t - \mu_2) + n(t), \quad -T < t < T. \tag{5-1}$$

We assume further that $\{n(t)\}$ is a stationary Gaussian process with zero mean, spectral density $K(\omega)$ and absolutely integrable covariance $k(\tau)$, statistically independent of $v(t) - n(t)$. We continue to assume that $\Phi(\omega) = \mathcal{F}[\phi(t)]$ is not zero for $|\omega| < \infty$.

The phenomenon of singular detection was noted in an early paper by Bussgang and Middleton [16] and also by Grenander [17]. Reference [14] contains a definitive collection of articles by Kallianpur and Oodaira [15], Yaglom [4], and Root [5]; from the latter paper, we shall need the

Theorem (Kelly, Reed, and Root): If the integral $I \triangleq \int_{-\infty}^{\infty} \frac{|\Phi(\omega)|^2}{K(\omega)} d\omega$ exists and is finite, detection is nonsingular. If $I = \infty$, detection is singular, at least when $T \to \infty$.

In the Medgyessy method of resolution the data $v(t)$ are filtered with a transfer function $1/\Phi(\omega)$, so that the output of the filter is $a_1\delta(t - \mu_1) + a_2\delta(t - \mu_2) + \nu(t)$. The noise term $\nu(t)$ is a Gaussian random variable of zero mean for each t. It is clear that if the variance of $\nu(t)$ is finite then perfect resolution results, for the noise term is insignificant compared to the impulses marking the wavelets' locations. Conversely, if $\operatorname{var} \nu(t) = \infty$ then perfect resolution is impossible, for the impulses are obscured by noise whose r.m.s. value is infinite at each t. Writing $\nu(t)$ as the convolution of $n(t)$ and the impulse response of $1/\Phi(\omega)$, and computing its variance yields

$$\operatorname{var} \nu(t) = \int_{-\infty}^{\infty} \frac{K(\omega)}{|\Phi(\omega)|^2} \frac{d\omega}{2\pi} \tag{5-2}$$

We observe that perfect resolution implies singular detection of $\phi(t)$ in the noise $\{n(t)\}$.

The modified method of quefrency alanysis estimates the power spectral density $P_{vv}(\omega)$ of the data $v(t)$. We shall suppose that the observation interval T is sufficiently large that the error in the estimate is negligibly small. Thus

$$P_{vv}(\omega) = K(\omega) + |\Phi(\omega)|^2 \frac{Q(\omega)}{T} \tag{5-3}$$

where $Q(\omega) = a_1^2 + a_2^2 + 2a_1a_2 \cos \omega(\mu_1 - \mu_2)$. On taking logarithms,

$$\ln P_{vv}(\omega) = \ln|\Phi(\omega)|^2 + \ln \frac{Q(\omega)}{T} + \ln\left[1 + \frac{K(\omega)}{|\Phi(\omega)|^2 \, Q(\omega)/T}\right] \tag{5-4}$$

Now the first term in equation (5-4) is subtracted away and the remainder subjected to inverse Fourier transformation. Using the inequalities $\ln(1 + x) \leqslant x$ and $Q(\omega) \geqslant (a_1 - a_2)^2$, the inverse transform of the last term in (5-4) is seen to be bounded by a constant times the integral in (5-2). Hence if perfect resolution is attainable by Medgyessy's method, it is attainable by quefrency alanysis also.

It is well known [5] that a nonsingular detection problem cannot be rendered singular by any amount of linear pre-filtering of the data. From the form of the integral (5-2) it is clear that the equivalent statement holds for the resolution problem.

VI. ILLUSTRATIVE EXAMPLE

Figure 1A shows a partially resolved triplet waveform typical of those produced by low resolution spectrometers. The waveform is composed of Gaussian wavelets having equal standard deviations of 0.4472 seconds possessing relative amplitudes of 1:2:1. The means of the wavelets are -1.0, 0.0, and 1.0 seconds respectively. The presence of the smaller wavelets is suggested by the inflections in the wings of the waveform.

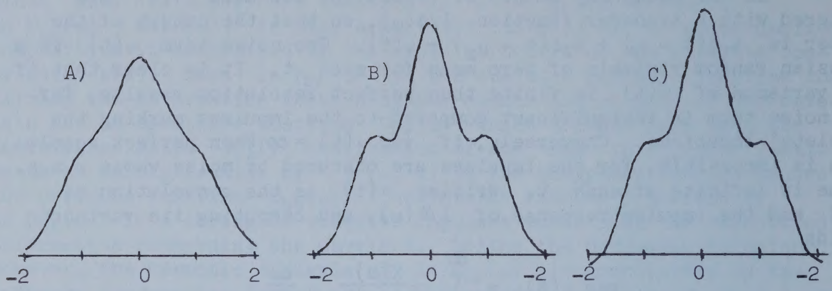


Figure 1. A) Simulated Partially Resolved Triplet Waveform.
B) Eddington Decomposition to Fourth Derivative.
C) Berger and van Cittert Decomposition to Three Terms of (3-21).

Figures 1B and 1C illustrate the application of the methods of Eddington and of Berger and van Cittert to improve the resolution of the curve shown in Figure 1A. In the method of Eddington we have employed only the second and fourth derivatives, and in the method of Berger and van Cittert, three terms of equation (3-21). Gladney [6] has observed that it is difficult to assign an accurate measure of the relative resolution enhancement furnished by the truncated versions of the various decomposition procedures. One measure of the enhancement is the ratio of the half-width of a single wavelet following application of a decomposition procedure to the half-width of the original wavelet. Under this criterion, the method of Eddington reduces the half-width by approximately 33%, while the method of Berger and van Cittert provides a reduction of approximately 20%. The reduction of half-width is neither the only nor the best criterion of improved resolution. The reduction of intensity in the tails of the wavelets is equally important. For the Gaussian wavelets this reduction is considerably better than the reduction in half-width would indicate.

REFERENCES

1) Lord Rayleigh (J. W. Strutt), Sci. Pap., vol. 1, 135 (1871)

2) P. Medgyessy, Decompositions of Superpositions of Distribution Functions, Plenum Press, New York (1961).

3) J. W. Tukey, Ref. 14, pp. 201-243.

4) A. M. Yaglom, Ref. 14, pp. 327-348.

5) W. L. Root, Ref. 14, pp. 292-315.

6) H. M. Gladney, J. Chem. Phys., vol. 40, 3135 (1964).

7) J. D. Morrison, J. Chem. Phys., vol. 39, 200 (1963).

8) H. Stone, J. Opt. Soc. America, vol. 52, 998 (1962).

9) R. N. Bracewell and J. A. Roberts, Australian J. Phys. vol. 7, 615, (1954).

10) A. Eddington, Mon. Nat. R. Astron. Soc., vol. 73, 359 (1913).

11) P. H. van Cittert and H. C. Berger, Z. Physik, vol. 81, 428 (1933).

12) H. Teicher, Annals Math. Stat., vol. 34, 1265 (1963).

13) H. Robbins, Annals Math. Stat., vol. 19, 360 (1948).

14) M. Rosenblatt (ed.), Proceedings of the Symposium on Time Series Analysis, Wiley, New York (1963).

15) G. Kallianpur and H. Oodaira, Ref. 14, pp. 279-291.

16) J. Bussgang and D. Middleton, IRE Trans., vol. IT-1, no. 5 (1955).

17) U. Grenander, Ark. Mat., vol. 1, pp. 195-277 (1950).

SOME APPLICATIONS OF REPRODUCING KERNEL HILBERT SPACES[f]

THOMAS KAILATH
Department of Electrical Engineering
Stanford University
Stanford, California

ABSTRACT

The representation of a second-order random process by a Hilbert space is now well known. So is the realization of a Hilbert space as the set, L^2, of (equivalence classes of) square-integrable functions. However, recently Parzen (Ann. Math. Stat., 1961) and Hajek (Czech. Math. J., 1962) have pointed out that a smaller space than L^2 is a more appropriate realization for many problems dealing with second-order processes. The functions of this smaller space, which is called a reproducing kernel Hilbert space (rkhs), are better behaved than the functions in L^2, for example, the functions in the rkhs are always continuous.

In this paper we first show how rkh spaces arise naturally in some signal detection problems. We then develop certain properties of such spaces and then apply them to the characterization of singular detection (Grenander, Arkiv für Mat., 1950) and stable detection (Root, AMS Appl. Math. Symposia, vol. XV, 1964) for deterministic signals in Gaussian noise. We also discuss a recent paper by Schwartz (IEEE-PGIT, July 1965).

SOME APPLICATIONS OF REPRODUCING KERNEL HILBERT SPACES

As Parzen has stressed in several papers [see, e.g., Parzen (1961), Parzen (1963)], reproducing kernel Hilbert spaces (rkhs) provide a natural coordinate free setting for many problems in communication theory. We shall illustrate this by showing how in a particular problem the introduction of the rkhs provides a coordinate free solution of the problem. We feel our illustration of this is somewhat simpler and more direct than Parzen's (cf. Parzen (1963)). We shall show the value of introducing the rkhs by simply translating some theorems derived by Aronszajn (1950) about rkhs into some results of communication-theoretic interest. In the last section we shall briefly mention some other applications.

1. A Coordinate-Free Description

We shall consider the following detection problem: Given an observation

---

[f] This work was supported by AFOSR Contract AF 49(638)1517 and Tri- Service Contract Nonr 225(83).

$x(t)$, $t \in T$, choose, in a way that will minimize the probability of error, between the hypotheses

$$\begin{aligned} h^{(1)} &: x(t) = m(t) + n(t) \\ h^{(0)} &: x(t) = n(t) \end{aligned} \tag{1}$$

where $n(t)$ is a sample from a zero mean, second-order Gaussian random process with covariance function

$$E[n(t)n(s)] = R(t,s) \quad , \quad t,s \in T \times T \tag{2}$$

and $m(t)$ is a waveform of known shape and of finite energy $[\int_0^T m^2(t)dt < \infty]$. A solution to this problem is well known--it was first given by Grenander (1950) and is also discussed by Davenport and Root (1958, Ch. XIV), Selin (1965, Ch. VII), Helstrom (1960, Ch. IV) and others. Their results are essentially the following: Decide $h^{(1)}$ only if (otherwise decide $h^{(0)}$)

$$\sum_1^\infty \frac{x_i m_i}{\lambda_i} > \text{a suitably chosen threshold } \delta \tag{3}$$

where x_i, m_i, λ_i are defined by

$$x_i = \int_T x(t)\psi_i(t)dt \ , \ m_i = \int_T m(t)\psi_i(t)dt \tag{4}$$

where $\{\psi_i(t)\}$ are the eigenfunctions of $R(t,s)$ over $T \times T$

$$\int_T R(t,s)\psi_i(s)ds = \lambda_i\psi_i(t) \ , \quad t \in T. \tag{5}$$

Furthermore, the probability of error depends monotonically upon a quantity D^2, called the detectability,

$$D^2 = \sum_1^\infty m_i^2/\lambda_i \ . \tag{6}$$

If $D^2 = \infty$, the error probability will be zero and the detection problem will be said to be singular. Such singular detection is not characteristic of physical problems and if it occurs we should conclude that the mathematical model of the physical problem has been overidealized. As an aid to mathematical model-making it would be desirable to have "simple" criteria for determining whether a mathematical model is singular or not. Some criteria of this sort have been given by Hajek (1960) and Kailath (1965) and will also be discussed below.

But first we note that the solution (3)-(6) is not "coordinate-free"-we can see this most easily as follows. Let $\{\psi_{wi}(t),\ i = 1,2,\ldots\}$ be the solutions

$$\int_T R(t,s)\psi_{wi}(s)w(s)ds = \lambda_{wi}\psi_{wi}(t)\ , \quad t\epsilon T \tag{7}$$

where $w(t)$, $t\epsilon T$, is a nonnegative (weight) function. Then we can show that if

$$\int_T\int_T R^2(t,s)w(t)w(s)dt\ ds < \infty\ ,\ w(t) \geq 0\ , \quad t\epsilon T \tag{8}$$

then the $\psi_{wi}(t)$ are orthonormal (w.r.t. $w(t)$),

$$\int_T \psi_{wi}(t)\psi_{wj}(t)w(t)dt = \delta_{ij} \tag{9}$$

and that Mercer's theorem holds,

$$R(t,s) = \sum_1^\infty \lambda_{wi}\psi_{wi}(t)\psi_{wi}(s) \tag{10}$$

where the series converges uniformly and absolutely. With these properties, one can establish that another solution to the detection problem (1) is: Decide $h^{(1)}$ only if

$$\sum \frac{x_{wi}m_{wi}}{\lambda_{wi}} > \text{a suitable threshold} \tag{11}$$

$$x_{wi} = \int_T x(t)\psi_{wi}(t)w(t)dt\ ,\ m_{wi} = \int_T m(t)\psi_{wi}(t)w(t)dt\ . \tag{12}$$

The error probability is essentially determined by a quantity

$$D_w^2 = \sum [m_{wi}]^2/\lambda_{wi}\ . \tag{13}$$

and detection will be nonsingular if and only if

$$\sum [m_{wi}]^2/\lambda_{wi} < \infty \tag{14}$$

These relations show the dependence of our solution upon the arbitrary (except for (8)) weight function and hence upon an arbitrary coordinate system $\psi_{wi}(t)$. It is natural to wonder what underlying property of $m(t)$, a property independent of $w(t)$, assures that condition (14) will be satisfied. We can determine this (coordinate free) property by using the notion of a reproducing kernel Hilbert space.

Definition: The reproducing kernel Hilbert space, $H(R,T)$, determined by the covariance function $R(t,s)$, $t, s \in T \times T$, is a Hilbert space with the following properties

(i) $R(\cdot,s) \in H(R,T)$ for every $s \in T$

(ii) If $m(\cdot) \in H(R,T)$ then $(m(\cdot), R(\cdot,s))_{H(R,T)} = m(s)$

where $(\ ,\ )_{H(R,T)}$ denotes the inner product in the space $H(R,T)$.

It is easy to show (see, e.g., Aronszajn (1950)) that the rkhs is unique, i.e., there is only one space $H(R,T)$ associated with a given covariance function $R(t,s)$, $t, s \in T \times T$. Using $H(R,T)$ we can rewrite the condition (14) in a coordinate free form:

$$\text{Detection is nonsingular if and only if } m(t) \in H(R,T) \tag{15}$$

The proof is simple. Let us make the space of functions $m(t) = \Sigma\, m_{wi}\psi_{wi}(t)$ that have $\Sigma[m_{wi}]^2/\lambda_{wi} < \infty$ into a Hilbert space, $\mathcal{H}$, with inner product

$$Q(a(t), b(t)) = \sum a_{wi}b_{wi}/\lambda_{wi} \tag{16}$$

where a_{wi}, b_{wi} are defined in the same way as m_{wi} (see (12)). This is the rkhs associated with $R(t,s)$ because

$$\text{(i)}\quad R(t,s) = \sum \lambda_{wi}\psi_{iw}(s)\psi_{iw}(t) \in \mathcal{H} \quad \text{since}$$

$$\sum [\lambda_{wi}\psi_{wi}(s)]^2/\lambda_{wi} = \sum \lambda_{wi}\psi_{wi}^2(s) = R(s,s) < \infty \tag{17}$$

$$\text{and}\quad \text{(ii)}\quad Q(R(t,s), m(t)) = \sum \lambda_{wi}\psi_{wi}(s)m_{wi}/\lambda_{wi} = \sum m_{wi}\psi_{wi}(s) = m(s) \tag{18}$$

By the uniqueness of the rkhs associated with $R(t,s)$, $t, s \in T \times T$, the rkhs $\mathcal{H}$ must be identical with (isomorphic to) $H(R,T)$. Therefore,

$$\sum \frac{m_{wi}^2}{\lambda_{wi}} < \infty \Leftrightarrow Q(m_w, m_w) = \|m_w\|^2_{H(R,T)} < \infty \tag{19}$$

which yields the statement (15). While the reformulation (15) is interesting because of its lack of dependence on $w(t)$ and $\psi_{wi}(t)$, it would be no more than this unless we could employ the intrinsic characterization (15) more usefully than the coordinatized characterization (14). We shall demonstrate the usefulness of (15) in Sec. 2. But first we shall also give an intrinsic characterization of the quantity

$$\sum x_{wi} m_{wi}/\lambda_{wi} = \Lambda , \text{ say.}$$

Our discussions might indicate that

$$\Lambda = (x(t), m(t))_{H(R,T)} \tag{20}$$

and this is formally correct. This is only formal because $x(t) = \Sigma x_{wi}\psi_{wi}(t)$ does not belong to $H(R,T)$. For if it have $\Sigma x_{wi}^2/\lambda_{wi} < \infty$. But it is easy to show that the x_{wi} defined by (12) are independent zero mean random variables with variances λ_{wi}. Therefore $E[\Sigma x_{wi}^2/\lambda_i] = \infty$ and hence, by a theorem of Kolmogorov [Doob (1953), p. 108]. $\Sigma x_{wi}^2/\lambda_i = \infty$ with probability one. Despite the formal nature of (20) it turns out to be quite useful.

2. Some Applications

In the last section we showed that a necessary and sufficient condition for nonsingular detection was that

$$\sum_1^\infty \frac{m_{wi}^2}{\lambda_{wi}} < \infty \tag{21}$$

This condition is not generally very useful since it requires computation of the eigenvalues and eigenfunctions of $R(t,s)$. In another paper (Kailath (1965)), we derived some more usable conditions for nonsingular detection. The derivations were based on the properties of the Karhunen-Loève expansion. However, they can also be obtained somewhat more directly by using the properties of the rkhs. As an example, we shall establish the following result derived differently in Kailath (1965). The present derivation will also suggest another useful result that was not obvious when the earlier Karhunen-Loève proof was obtained. The result is:

The detection of $m(t)$, $t\epsilon T$, is singular[†] if all extensions $m_\infty(t)$, $-\infty < t < \infty$, $[m_\infty(t) = m(t), t\epsilon T]$ are singular over $(-\infty,\infty)$. The detection of $m(t)$, $t\epsilon T$, is nonsingular if at least one extension is nonsingular over $(-\infty,\infty)$ (22)

For a proof we recall that in Sec. 1 we showed that condition (21) was equivalent to:

[†] We use this term as shorthand for "m(t) will be capable of singular detection."

Detection is nonsingular over T if and only if $m(t) \in H(R,T)$. (23)[*] We now have to assume that we can extend the covariance function $R(t,s)$, $t, s \in T \times T$ to a covariance function, $R_\infty(t,s)$ defined over the whole plane $(-\infty < t, s < \infty)$. In many problems there will be a "natural" extension R_∞. Let $H(R_\infty)$ be the rkhs associated with R_∞. Then it is easy to see that our result (22) is a consequence of the following rkhs theorem (Aronszajn, 1950, p. 351):

If R and R_∞ are related as above then the elements of $H(R,T)$, say f_T, are restrictions[†] to T of functions f_∞ belonging to $H(R_\infty)$. Furthermore,

$$\|f_T\|_{H(R,T)} = \min_{f_\infty} \|f_\infty\|_{H(R_\infty)} \qquad (24)$$

where $f_\infty(t)$, $-\infty < t < \infty$, are the functions in $H(R_\infty)$ whose restriction to T is just f_T.

The result (22) will not be much use unless it is easier to show that an extension, m_∞, of $m(t)$, $t \in T$, does or does not belong to $H(R_\infty)$ than it is to determine whether or not $m(t)$, $t \in T$, belongs to $H(R,T)$. Fortunately, as is shown in Kailath (1965) in many cases it is possible to do so.

For example if the noise is stationary with spectral density $S_\infty(f)$ we can show that

$$\|m_\infty\|^2_{H(R_\infty)} = \int_{-\infty}^{\infty} |M_\infty(f)|^2/S_\infty(f)df . \qquad (25)$$

This involves only the comparatively easy calculation of Fourier transforms while evaluating $\|m\|^2_{H(R,T)}$ involves computing eigenvalues and eigenfunctions or solving certain integral equations.

At this point we may also notice another useful application of Aronszajn's theorem (24). Since it is often easier to calculate $\|m_\infty\|_{H(R_\infty)}$ than it is to calculate $D^2 = \|m_T\|_{H(R,T)}$, (24) shows that we can find simple upper bounds to D^2 by computing $\|m_\infty\|_{H(R_\infty)}$ for some nonsingular extension m_∞. If we can carry out the minimization in (23), we can determine D^2 exactly, but, even if this cannot be done, a few "educated" trials should give a reasonably good estimate of D^2.

We may note that this useful possibility of approximating D^2 was directly suggested by the rkhs theorem--it can also be shown using the Karhunen-Loève approach of Kailath (1965), but in that framework it is not so obvious.

---

[*] This result is true even if T is infinite, $T = (-\infty,\infty)$.

[†] That is, $f_T(t) = f_\infty(t)$, $t \in T$, $f_T(t) = 0$, elsewhere.

The fact that $m(t), t\epsilon T$ must belong to $H(R,T)$ for nonsingular detection also has some other fairly obvious conclusions. We note that $H(R,T)$ is a subspace of $L^2(T)$ [the Hilbert space of square-integrable functions on T]. Therefore, for $m(t)$ to belong to $H(R,T)$ it is not sufficient for $m(t)$ to be square-integrable,--$m(t)$ must have additional smoothness properties. And in fact, it is fairly easy to prove [see, e.g., Parzen (1959)] that the elements of $H(R,T)$ are always continuous and that if $\partial^{2m}R(t,s)/\partial t^m \partial s^m$ exists then the elements are at least m-times differentiable. Along the same lines, we can prove that if the covariance $R(t), t\epsilon T$, is the restriction to T of the covariance of a process with spectral density (RC noise)

$$S(f) = \frac{1}{\alpha^2 + 4\pi^2 f^2} \tag{26}$$

then

$$H(R,T) = \left\{\begin{array}{l}\text{space of all functions } f(t),\ t\epsilon T \text{ with} \\ \text{square integrable derivatives.}\end{array}\right\} \tag{27}$$

We shall conclude this paper with two applications of the results just mentioned. The first deals with what Root (1964) calls the stability of the detection problem (1)-(2).

Root defines the detection problem to be stable if slight errors in the assumed forms of signal $m(t)$ and covariance $R(t,s)$ do not result in a large change in the error probabilities. He notes, for example, the following. Suppose the true signal differs from the nominal signal $m(t)$ by $\epsilon(t)$ where $\int_T \epsilon^2(t)dt < \infty$. If, using the definitions in (4)-(5), $\Sigma\, m_i^2/\lambda_i^2 = \infty$, then there exists an $\epsilon(t)$, of arbitrarily small energy, such that the probability of error is zero. [It is assumed that when $\epsilon(t) = 0$ the error probability is not zero.] This result lends special importance to the situations where $\Sigma\, m_i^2/\lambda_i^2 = \infty$. But since this condition is not satisfied in many detection problems (see, e.g., Kailath (1965b) this instability result is rather disturbing. Such extreme sensitivity is not characteristic of physical problems and there must be something unrealistic in the mathematical model. The properties of the rkhs that we have just discussed help to locate the source of the difficulty, which is rather simple. For nonsingular detection, we have seen that the signal must belong to the rkhs, $H(R,T)$ of $R(t,s)$. If both the true signal and the nominal signal belong to $H(R,T)$ then so must the error $\epsilon(t)$. But in establishing Root's instability result it is necessary to allow $\epsilon(t)$ to be any function of finite energy and such a function may not be smooth enough to yield nonsingular detection. This is the difficulty with the model--if $\epsilon(t)$ is in $H(R,T)$, which is a reasonable restriction, then there is no instability.

Finally we mention a recent paper by Schwartz (1965), which studies the detection of a signal $m(t)$, $t \epsilon T$, in a noise with spectral density given by (26). Schwartz computes the quantity Λ, of Eq. 20, as[†] (in our notation)

$$\Lambda = \int_0^T x(t)[\alpha^2 m(t) - \ddot{m}(t)]dt + [\alpha m(0) - \dot{m}(0)]x(0) + [\alpha m(T) + \dot{m}(T)]x(T) \qquad (28)$$

where $\ddot{m}(t) = d^2m(t)/dt^2$, $\dot{m}(0) = dm(t)/dt|_{t=0}$. However, Schwartz obtains this result by a rather detailed argument that requires the signal $m(t)$ to have a *continuous third derivative*. Schwartz also points out that the result (28) can be obtained more rapidly though presumably less rigorously, by an integral equation approach discussed, for example, in Helstrom (1960), Ch. IV).

Our rkhs result (27) shows that for nonsingular detection $m(t)$ does not need to be as smooth as Schwartz assumes--viz., it need only have a square integrable derivative. With this assumption, some computation yields

$$\Lambda = \int_0^T [\dot{x}(t) + \alpha x(t)][\dot{m}(t) + \alpha m(t)]dt + 2\alpha x(0)m(0) \qquad (29)$$

This form may not be desirable because it involves the (nonexistent) derivative $\dot{x}(t)$,[††] but an integration by parts, assuming that $m(t)$ has *two square-integrable derivatives*, will yield the form (28). Moreover, we can establish as follows that (28) is "rigorously" true--obtain Λ by any means, rigorous or otherwise, an expression for Λ, but then check that it can be written $\Lambda = (x(t), m(t))_{H(R,T)}$. If this holds (and it does for (28)), the formula for Λ will be correct.

[†] We are simplifying Schwartz's solution--his Eq. 18--a bit by setting, in his notation $S^{(2)}(t) \equiv 0$, and neglecting the terms that do not depend upon the received (noisy) signal $Y(t)$. We can obtain his form by setting $m(t) = S^{(1)}(t) - S^{(2)}(t)$.

[††] I. Selin pointed out in a private discussion that while $\dot{x}(t)$ does not exist for a noise (26), and therefore cannot be computed, the quantity $\dot{x}(t) + \alpha x(t)$ can be computed reasonably accurately.

3. Concluding Remarks

We have illustrated, with a few examples, the value of introducing the notion of a rkhs in certain detection problems. Similar advantages can be illustrated in several other detection and estimation problems.

REFERENCES

1. Aronszajn, N., "Theory of Reproducing Kernels," Trans. Amer. Math. Soc., vol. 63, no. 3, pp. 337-404, May 1950.
2. Davenport, W. and Root, W. L., An Introduction to the Theory of Random Signals and Noise, McGraw-Hill Book Co., New York, 1958.
3. Doob, J. L., Stochastic Processes, John Wiley and Sons, Inc., New York, 1953.
4. Grenander, U., "Stochastic Processes and Statistical Inference," Arkiv für Mat., vol. 1, pp. 195-277, 1950.
5. Hajek, J., "On Linear Statistical Problems in Stochastic Processes," Czech. Math. J., vol. 12, pp. 404-444, 1962.
6. Helstrom, C. W., Statistical Theory of Signal Detection, Pergamon Press, 1960.
7. Kailath, T., "Some Results on Singular Detection, submitted to Information and Control.
8. Kailath, T., "The Detection of Known Signals in Colored Gaussian Noise," Proc. NEC, Chicago, Oct. 1965.
9. Parzen, E., "Statistical Inference on Time Series by Hilbert Space Methods," Tech. Rept. No. 23, Appl. Math. and Stat. Lab., Stanford Univ., 1959.
10. Parzen, E., "A New Approach to the Synthesis of Optimal Smoothing and Prediction Systems," Ch. V of Math. Optimization Techniques, ed. by R. Bellman, Univ. of Calif. Press, 1963.
11. Root, W. L., "Stability in Signal Detection," Proc. Symposia in Applied Math., vol XVI, Amer. Math. Soc., Rhode Island, 1964.
12. Schwartz, M. I., "On the Detection of Known Binary Signals in Gaussian Noise of Exponential Covariance," Trans. IEEE-PGIT, vol. IT-11 pp. 330-335, July 1965.

SOME CRITERIA FOR FREQUENCY DOMAIN DESIGN OF SIGNALS IMBEDDED IN GAUSSIAN NOISE

MARVIN K. SIMON and LUDWIK KURZ
New York University
Bronx, New York 10453

ABSTRACT

The design of time-limited binary signals imbedded in colored gaussian noise resulting from passing white noise through a channel of known transfer function is investigated. It is shown that, for a class of channels whose magnitude characteristic is expressible as the ratio of polynomials with the numerator divisible by and of lower or the same degree as the denominator, the output noise kernel (correlation function) and its inverse will have Fourier transforms which are reciprocals of each other. The above assumption results in an expression for probability of error which is independent of the received waveform and depends solely on the transmitter signal-to-noise ratio. The additional freedom of design afforded by the above enables one to impose further restrictions in the form of energy concentration in the frequency domain on the received signal without deterioration of system performance. Some such bandwidth constraints and their associated optimum signal designs are considered in this paper.

1. INTRODUCTION

Consider a special case of a generalized communication channel model illustrated in Figure 1, namely:

1) $|H(\omega)|^2 = \frac{N(\omega^2)}{D(\omega^2)}$ where $N(\omega^2)$ and $D(\omega^2)$ are rational polynomials and $N(\omega^2)$ is of lower or the same degree as $D(\omega^2)$.
2) $n_1(t)$ is a gaussian noise process with spectral density $S_{n_1}(\omega) = \frac{P(\omega^2)}{Q(\omega^2)}$ where $P(\omega^2)$ and $Q(\omega^2)$ are rational polynomials and $P(\omega^2)$ is of lower or the same degree as $Q(\omega^2)$.
3) $n_2(t) = 0$
4) $s(t)$ and $x(t)$ are both timelimited to $0 \leq t \leq T$.

Results obtainable under the above assumptions are applicable to a broad class of systems in which the channel is described by a lumped element equivalent, the noise introduced at the receiver is negligible with respect to that introduced before the receiver, and no intersymbol interference is present (i.e. time-limited sequences).

The total output signal $y(t) = s(t) + n(t)$ upon which a decision (continuous or discrete) is to be made has the following properties:

1) $S(\omega) = H(\omega)X(\omega)$ where $S(\omega)$ and $X(\omega)$ are the Fourier Transforms of $s(t)$ and $x(t)$ respectively.
2) $S_n(\omega) = |H(\omega)|^2 S_{n_1}(\omega)$ with corresponding autocorrelation function $R_n(\tau)$.

For binary continuous signalling in colored gaussian noise, an optimum decision based on the maximum likelihood principle (an equal cost equal a priori Bayes criterion) results in a binary symmetric channel with probability of error

$$p_e = \tfrac{1}{2}\operatorname{erfc}\left[\frac{\displaystyle\iint_0^T s(t)s(p)r_n(p-t)\,dt\,dp}{2}\right]^{\frac{1}{2}}$$

where the inverse noise kernel $r_n(p-t)$ is defined by

$$\int_0^T R_n(p-t)r_n(t-u)dt = \delta(p-u) \qquad \begin{matrix} 0 < p < T \\ 0 \le u \le T \end{matrix} \qquad (1)$$

It is of particular interest (as will be seen shortly) to investigate a class of kernels $R_n(\tau)$ which have the property that

$$\mathcal{F}\{r_n(\tau)\} = \frac{1}{\mathcal{F}\{R_n(\tau)\}}$$

where $\mathcal{F}$ represents Fourier Transform.

Let the above integral equation defining the inverse kernel be satisfied in addition for $0 \le u \le T$ and all p, $(-\infty \le p \le \infty)$. Then, substituting $\tau = p-u$, and taking the Fourier transform of both sides of (1),

$$\int_{-\infty}^{\infty} e^{-j\omega\tau}\,d\tau \int_0^T R_n(\tau-(t-u))r_n(t-u)dt = 1 \qquad -\infty \le \tau \le \infty$$

Integrating first with respect to τ,

$$S_n(\omega)e^{j\omega u}\int_0^T r_n(t-u)e^{-j\omega t}\,dt = 1 \qquad 0 \le u \le T \qquad (2)$$

If $r_n(t)$ is of finite support and such that a time shift of u for any **u** in the interval (0,T), does not shift any part of the function outside the interval (0,T), then,

$$\int_0^T r_n(t-u)e^{-j\omega t}\,dt = \mathcal{F}\{r_n(\tau)e^{-j\omega u}$$

which together with (2) arrives at the desired result,

$$S_n(\omega)\mathcal{F}\{r_n(\tau)\} = 1 \qquad (3)$$

A set of functions which satisfy the above requirements on $r_n(t)$ is the impulse function and all its derivatives. Thus, if $S_n(\omega)$ is expressible as a ratio of polynomials $A(\omega^2)/B(\omega^2)$ where $A(\omega^2)$ is of smaller or the same degree as $B(\omega^2)$, the corresponding kernel and its inverse will have Fourier Transforms which are reciprocal. (It is also implied that $B(\omega^2)$ is divisible by $A(\omega^2)$.) For the class of channels and input noise processes being considered, the above is automatically satisfied.

Returning now to the expression for probability of error,

$$\int_0^T\!\!\int s(t)s(p)r_n(p-t)\,dt\,dp = \int_0^T\!\!\int s(t)s(p)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{j\omega(p-t)}}{|H(\omega)|^2\ S_{n1}(\omega)}\,d\omega\right]dt\,dp$$

Factoring $S_{n1}(\omega)$ into $S_{n1}(\omega) = S_{n1}(\omega)^+ S_{n1}(\omega)^-$ where $S_{n1}{}^+(\omega)$ has all its poles and zeros in the upper half ω-plane and $S_{n1}{}^-(\omega)$ has all its poles and zeros in the lower half ω-plane, and noting that $S_{n1}{}^-(\omega) = (S_{n1}{}^+(\omega))^*$, where the star denotes complex conjugate,

$$\int_0^T\!\!\int s(t)s(p)r_n(p-t)\,dt\,dp = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\frac{X(\omega)}{S_{n1}{}^+(\omega)}\right|^2 d\omega \tag{4}$$

which is a measure of signal to noise ratio at the input. Indeed, for a white noise input with autocorrelation function $R_{n1}(\tau) = N_o/2\ \delta(\tau)$, the above expression reduces to

$$\int_0^T\!\!\int s(t)s(p)r_n(p-t)\,dt\,dp = \frac{2\int_0^T x^2(t)\,dt}{N_o} \tag{5}$$

For a fixed average power at the transmitter, S^2, the probability of error is then independent of the output waveform and depends solely on the input signal to noise ratio as illustrated by

$$p_e = \tfrac{1}{2}\text{erfc}\left[\sqrt{\frac{2\ S^2T}{N_o}}\right] \tag{6}$$

The additional freedom of design afforded by the above results (for white noise input) enables one to impose further restrictions on s(t) without deterioration of system performance in terms of error probability.

From the standpoint of minimizing interchannel interference, it is desirable that s(t) have some degree of energy concentration in the frequency domain. The discussion which follows presents several practical criteria pertaining to reduction of frequency spillover.

2. GENERALIZED VARIANCE CRITERION

As a measure of spread or duration of a signal, one often uses the second moment of its energy spectrum about a suitably chosen point, e.g. the position of its maximum or its center of gravity. Since the frequency energy spectrum of a real time function is an even function of ω about the origin, the variance of $|S(\omega)|^2$ is then defined as,

$$\epsilon_\omega^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\omega^2\ |S(\omega)|^2\ d\omega\ . \tag{7}$$

For frequency energy spectra which are not strictly low pass in nature, one should consider a more general criterion of this type, namely,

$$\epsilon_\omega^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(\omega)\ |S(\omega)|^2 d\omega \tag{8}$$

where the weighting function $g(\omega)$ is assumed to be an even function of ω about the origin and has a Taylor Series expansion in even powers of ω, with coefficients c_i, $i = 0,2,4, \ldots$.

Using the moment theorem for Fourier transforms and Parseval's Theorem,

$$\epsilon_\omega^2 = \int_0^T \left[\sum_{-\infty}^{\infty} c_{2i} (s^{(i)}(t))^2 \right] dt \tag{9}$$

where the superscript notation indicates i^{th} order differentiation. Since the input is average power limited, the design of $s(t)$ may be formulated in terms of an isoperimetric problem in the calculus of variations, namely: minimize ϵ_ω^2 subject to the constraint

$$\iint_0^T s(t)s(p)r_n(p-t)dt\, dp = \frac{2\, S^2 T}{N_o} \tag{10}$$

If

$$\gamma = \frac{\epsilon_\omega^2}{S^2 T} \tag{11}$$

then the corresponding Euler's (also known as Euler-Lagrange's equation) equation becomes

$$\int_0^T s_k(p) r_n(p-t) dp = \frac{\sum_{i=0}^{\infty} (-1)^i c_{2i} s_k^{(2i)}(t)}{\gamma_k \frac{N_o}{2}} \qquad 0 \leq t \leq T \tag{12}$$

Having derived the general result, its application to a particular example is now considered with emphasis on the tractability of the resulting signal design.

Let $H(\omega)$ be a single pole low pass filter, i.e.

$$H(\omega) = \frac{1}{1 + j\omega/\Omega}$$

Then, for a white noise input as previously assumed, the output noise is a Markovian process with autocorrelation function

$$R_n(\tau) = \frac{N_o \Omega}{4} e^{-\Omega(\tau)} \text{ and spectral density } S_n(\omega) = \frac{(N_o/2)\,\Omega^2}{\Omega^2 + \omega^2} \;.$$

The inverse noise kernel $r_n(\tau)$ is determined from the reciprocal of $S_n(\omega)$ as

$$r_n(\tau) = \frac{\delta(\tau)}{N_o/2} - \frac{\delta^{(2)}(\tau)}{(N_o/2)\,\Omega^2} \;.$$

Since the channel proposed is a low pass structure, the variance criterion for frequency energy concentration is quite suitable. Thus, letting $c_2 = 1$ and all other $c_{2i} = 0$, Euler's equation becomes for this particular example,

$$\frac{d^2 s_k(t)}{dt^2} + \left(\frac{\gamma_k \omega_o^2}{\Omega^2 - \gamma_k} \right) s_k(t) = 0 \qquad 0 \leq t \leq T$$

which has a general solution of the form

$$s_k(t) = A_k \sin \sqrt{\frac{\gamma_k \Omega^2}{\Omega^2 - \gamma_k}}\, t + B_k \cos \sqrt{\frac{\gamma_k \Omega^2}{\Omega^2 - \gamma_k}} \qquad 0 \le t \le T$$

Since the denominator of $H(\omega)$ represents proportional plus derivative control, any jump discontinuity in $s(t)$(in particular at the end-points $t = 0,T$) would require an impulse function in $x(t)$ thereby violating the requirement of fixed finite input energy. It is therefore necessary to impose the end conditions $s(0) = s(T) = 0$, from which we conclude $B_k = 0$ and

$$\left(\frac{\gamma_k \Omega^2}{\Omega^2 - \gamma_k} \right)^{\frac{1}{2}} T = k\pi \qquad k = 1,2,\ldots$$

The smallest of the set of eigenvalues

$$\gamma_k = \frac{(k\pi/T)^2}{1 + (k\pi/\Omega T)^2} \tag{13}$$

occurs for k = 1, with corresponding eigenfunction

$$s_1(t) = A_1 \sin \frac{\pi}{T} t \qquad 0 \le t \le T \tag{14}$$

Finally, for an average power constraint at the transmitter, S^2, the constant A_1 may be evaluated as

$$A_1 = \sqrt{\frac{2\, S^2 T}{T[1 + (\pi/\Omega T)^2]}} \tag{15}$$

which completes the optimum received signal design. It is interesting to note that this signal is also the time limited output signal of maximum energy for a given input energy, $E^2 = S^2 T$.[1] This property is desirable when a decision is to be made in the presence of white noise only.

3. MAXIMUM CONCENTRATION OF ENERGY IN A GIVEN FREQUENCY BAND

A second criterion for concentration of energy in the frequency domain which has been used successfully in recent applications,[2,3,4] is to maximize the energy in a given frequency band, i.e.,

$$\beta^2 E_0^2 = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} |S(\omega)|^2 \, d\omega \tag{16}$$

where E_0^2 is the total output energy.

Transforming the above to the time domain and changing the interval of definition of $s(t)$ to $(-T/2, T/2)$ to obtain symmetrical eigenfunctions,

$$\beta^2 E_0^2 = \iint_{-T/2}^{T/2} s(p)s(t)\, \frac{\sin \Omega(p-t)}{\pi(p-t)}\, dt\, dp \tag{17}$$

where $\frac{\sin \Omega(p-t)}{\pi(p-t)}$ represents the "brick wall" kernel.

The isoperimetric problem for designing $s(t)$ according to the above criterion is described by: maximize $\beta^2 E_0^2$ subject to the constraint (10). Defining $\eta = \frac{\beta^2 E_0^2}{(2S^2T/N_0)}$, the associated Euler's equation becomes

$$\int_{-T/2}^{T/2} s(p)\left[r_n(p-t) - \frac{\sin \Omega(p-t)}{\eta\pi(p-t)}\right] dp = 0 \tag{18}$$

Of particular interest is the application of the above to the example just considered in the previous sub-section, whereupon

$$\int_{-T/2}^{T/2} s_k(p)\, \frac{\sin \Omega(p-t)}{\pi(p-t)}\, dp = \frac{\eta_k}{N_0/2}\left[s_k(t) - \frac{1}{\Omega^2}\, s_k^{(2)}(t)\right] \tag{19}$$

$$-T/2 \leq t \leq T/2$$

Expanding the "brick wall" kernel using Mercer's Theorem into a uniformly convergent series of its Fredholm eigenfunctions, i.e.

$$\frac{\sin \Omega(p-t)}{\pi(p-t)} = \sum_n \lambda_n \frac{\psi_n(p)}{\sqrt{\lambda_n}} \frac{\psi_n(t)}{\sqrt{\lambda_n}} \tag{20}$$

where $\psi_n(t)$ is the n^{th} order prolate spheroidal function and λ_n is the n^{th} order eigenvalue[5], and also expanding $s_k(t)$ into an infinite series of the functions $\psi_n(t)$ with coefficients c_{kn}, one obtains

$$\int_{-T/2}^{T/2} \sum_n c_{kn}\psi_n(p) \sum_\ell \psi_\ell(p)\psi_\ell(t)\,dp = \frac{\eta_k}{N_0/2}\left[\sum_n c_{kn}\psi_n(t) - \frac{1}{\Omega^2}\sum_r c_{kr}\psi_r^{(2)}(t)\right] \tag{21}$$

$$-T/2 \leq t \leq T/2$$

Using the properties of the eigenfunctions $\{\psi_k(t)\}$, letting $\rho_k = \frac{\eta_k}{N_0/2}$ for simplicity of notation, and after multiplying both sides of the equation by $\psi_i(t)$ and integrating between $-T/2$ and $T/2$,

$$c_{ki}\lambda_i(\lambda_i - \rho_k) + \frac{\rho_k}{\Omega^2}\sum_r c_{kr} d_{ir} = 0 \qquad i = 0,1,2,\ \ldots \tag{22}$$

where

$$d_{ir} = \int_{-T/2}^{T/2} \psi_i(t)\psi_r^{(2)}(t)\ dt. \tag{23}$$

This is a set of homogeneous equations and thus has a solution for c_{ki}, $i = 0,1,2,\ \ldots$, iff the determinant of its coefficients vanishes. Truncating the orthogonal series expansion for $s_k(t)$ to N terms, N of the eigenvalues ρ_k may be approximately determined from

$$\det \begin{bmatrix} \lambda_o(\lambda_o-\rho_k)+\frac{\rho_k d_{oo}}{\Omega^2} & \frac{\rho_k d_{o1}}{\Omega^2} & \frac{\rho_k d_{o2}}{\Omega^2} & \cdots & \frac{\rho_k d_{oN}}{\Omega^2} \\ \frac{\rho_k d_{1o}}{\Omega^2} & \lambda_1(\lambda_1-\rho_k)+\frac{\rho_k d_{11}}{\Omega^2} & \frac{\rho_k d_{12}}{\Omega^2} & \cdots & \frac{\rho_k d_{1N}}{\Omega^2} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \frac{\rho_k d_{No}}{\Omega^2} & \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot & & & \lambda_N(\lambda_N-\rho_k)+\frac{\rho_k d_{NN}}{\Omega^2} \end{bmatrix} = 0$$

A set of N-dimensional eigenvectors $s_k(t)$ is then constructed by solving the previous set of homogeneous equations for each of the N eigenvalues just determined. The optimum received signal $s(t)$ is derived by combining a minimum set of eigenvectors corresponding to the largest eigenvalues necessary to satisfy the end conditions.

Expressions for the coefficients d_{ir} for $i \neq r$ in terms of known or tabulated quantities can be derived[6] with the results given below:

$$d_{ir} = \int_{-T/2}^{T/2} \psi_i(t)\psi_r^{(2)}(t)\,dt = \frac{2\lambda_i}{\lambda_i - \lambda_r}\left[\psi_r(T/2)\psi_i^{(1)}(T/2) - \psi_r^{(1)}(T/2)\psi_i(T/2)\right] \qquad \text{for } |r-i| \text{ even} \qquad (24a)$$

$$= \quad 0 \qquad \text{for } |r-i| \text{ odd} \qquad (24b)$$

Evaluation of the coefficients d_{ii} may be performed using numerical integration techniques, since a closed form solution is unable to be found with the properties of prolate spheroidal functions presently available in the literature. Also note that $d_{ir}/d_{ri} = \lambda_i/\lambda_r$ and thus the matrix from which the eigenvalues ρ_k are determined is not symmetric.

4. MODIFICATION OF THE "BRICK WALL" KERNEL FOR FREQUENCY ENERGY CONVERSION

The final criterion on frequency concentration to be discussed is a modification of the former, replacing the "brick wall" kernel by a "softer" version. A kernel whose Fourier Transform is a single pole low pass filter is considered thus eliminating the necessity of dealing with the prolate spheroidal functions about whose properties relatively little is known. Assuming a "soft wall" characteristic whose inverse transform is identical with the noise kernel

$$R_n(\tau) = \frac{\Omega}{2}\, e^{-\Omega(\tau)} \quad ,$$

one obtains an Euler equation for the particular channel proposed,

$$\int_{-T/2}^{T/2} s_k(v)\,\frac{\Omega}{2}\, e^{-\Omega|v-t|}\,dp = \tau_k\left[s_k(t) - \frac{1}{\Omega^2}\, s_k^{(2)}(t)\right] \qquad -T/2 \leq t \leq T/2 \qquad (25)$$

where

$$\tau_k = \frac{\iint_{-T/2}^{T/2} s_k(v)s_k(t)\,\frac{\Omega}{2}\,e^{-\Omega|v-t|}\,dt\,dv}{S^2T}$$

Since the left side of the above is a convolution integral,

$$\Omega^2 s_k(t) = \tau_k\left[\Omega^2-p^2\right]\left[s_k(t) - \frac{1}{\Omega^2}\,s_k^{(2)}(t)\right] \qquad -T/2 \le t \le T/2 \qquad (26)$$

where p represents the differential operator d/dt. Expanding the above, $s_k(t)$ must satisfy the differential equation,

$$s_k^{(4)}(t) - 2\,\Omega^2 s_k^{(2)}(t) + \Omega^4\left(1 - \frac{1}{\tau_k}\right)s_k(t) = 0 \qquad -T/2 \le t \le T/2 \quad (27)$$

The solution for the eigenfunctions $\{s_k(t)\}$ is of the form

$$s_k(t) = A_k \sin b_k t + B_k \cos b_k t \qquad -T/2 \le t \le T/2 \quad k=1,2,\ldots \qquad (28)$$

Substitution of (28) in Euler's equation (in both integral and differential form) yields

$$\tau_k = \frac{1}{1 + \frac{2b_k^2}{\Omega^2} + \frac{b_k^4}{\Omega^2}} \qquad (29)$$

and

$$\begin{aligned} b_k \tan b_k\, T/2 &= \Omega \text{ if } s_k(t) = B_k \cos b_k t \\ b_k \cot b_k\, T/2 &= \Omega \text{ if } s_k(t) = A_k \sin b_k t \end{aligned} \qquad (30)$$

Using τ_k as a measure of frequency concentration, the optimum signal s(t) is formed from a linear combination of the two eigenvectors $s_1(t)$ and $s_2(t)$ corresponding to the two largest eigenvalues τ_1 and τ_2 with appropriate coefficients chosen to satisfy the end conditions and fixed input energy constraint.

In conclusion, one notes that the last two criteria suffer from the fact that for binary bipolar signalling, more than one eigenfunction is necessary to satisfy the required end conditions on s(t). If it is desired to build up higher order alphabets by an orthogonal digit code based on these eigenfunctions, each digit must consist of at least two eigenfunctions with coefficients adjusted to meet the signal end conditions and digit orthogonality.

REFERENCES

1. M. I. Schwartz. "Binary Signal and Receiver Design for Linear Time Invariant Channels," a thesis, New York University, May, 1964. An appendix in the original draft (not available in the final version).
2. B. R. Saltzberg and L. Kurz. "Design of Bandlimited Signals for Binary Communication Using Simple Correlation Detection," B. S. T. J., 44 (1965), pp. 235-252.

3. H. Landau, and H.O. Pollak. "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - II," *B.S.T.J.*, 40 (1961), pp. 65-84.
4. M.K. Simon and L. Kurz. "A Method of Signal Design for Gaussian Noise and Intersymbol Interference Immunity using Maximum Likelihood Detection," *Tech. Rpt. 400-110*, New York University, (1965). (Also available in part as a paper in *Symposium on Signal Transmission and Processing*, IEEE, Group on Circuit Theory, Monograph 4C9, May 1965, pp. 62-69.)
5. D. Slepian and H.O. Pollak. "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-I," *B.S.T.J.*, 40 (1961), pp. 43-63.
6. M.K. Simon and L. Kurz. "Optimal Processing and Design of Digital Signals Perturbed by Gaussian and Nongaussian Noise," *Tech. Rept. 400-121*, New York University, (1965).

ACKNOWLEDGMENT

The research reported herein was supported in part by the National Science Foundation under grant no. NSF-GP-2942.

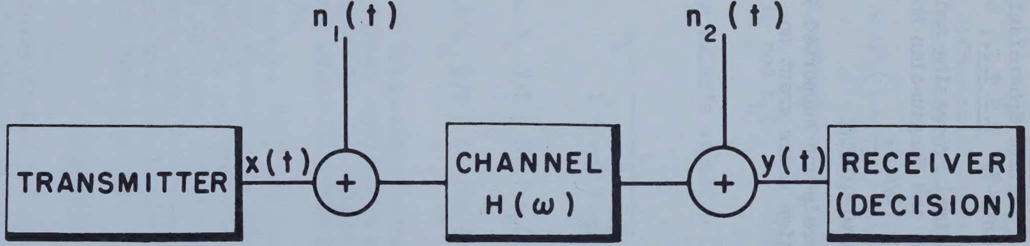

n_1(t)
n_2(t)
TRANSMITTER
x(t)
+
CHANNEL
H(ω)
+
y(t)
RECEIVER
(DECISION)

FIG. 1

SYSTEM MODEL

OPTIMAL OPEN-LOOP CONTROL OF LINEAR SYSTEMS WITH TIME LAG

THOMAS E. MUELLER
Department of Electrical Engineering
University of Texas
Austin, Texas

ABSTRACT

Results are given which permit a numerical open-loop solution of the optimal regular problem with a finite optimization interval for systems governed by linear differential-difference equations. An iterative algorithm is derived from the necessary and sufficient conditions for optimality which requires neither the choice of an initial control nor the choice of a convergence parameter.

INTRODUCTION

Some progress has been made in recent years in obtaining a solution to the optimal regulator problem with a system of linear differential-difference equations as constraints. N. N. Krasovskii has investigated this problem in [1] and [2] and has derived the form of the closed-loop controller. However, the solution to his equations for the controller functional is very difficult and requires various approximations to obtain a practical numerical solution. In addition, the controller in closed-loop form is impractical to implement physically. In [3], Kramer has found the solution to a special case of the problem treated in this paper where only the final states and and integral of the controls appear in the performance index. In this paper, an approach based on the gradient technique is used to derive a computer algorithm which results in a decreasing sequence of performance indices and requires neither the choice of a convergence parameter nor the choice of a starting control.

STATEMENT OF THE PROBLEM

A vector set of linear differential-difference equations is given by

$$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{x}(t-\tau) + C(t)\underline{u}(t) \ , \qquad (1)$$

defined for the interval $t_0 < t \leq t_1$, where $\underline{x}(t)$ is a real valued n-dimensional vector, $\underline{u}(t)$ is a real valued r-dimensional control vector, A(t) and B(t) are given real valued n × n matrices of continuous functions of t on $t_0 \leq t \leq t_1$, C(t) is a given real valued n × r matrix of continuous functions on t on the same interval, t_0 and t_1 are real fixed and finite constants, and τ is a real positive constant. An initial function for these equations is given by

$$\underline{x}(t) = \underline{g}(t) \qquad (2)$$

defined for $t_0 - \tau \leq t \leq t_0$. A performance index J is defined by

$$J = \frac{1}{2}\,\underline{x}^T(t_1)P\underline{x}(t_1) + \int_{t_0}^{t_1} \{ \frac{1}{2}\,\underline{x}^T(t)Q(t)\underline{x}(t) + \frac{1}{2}\,\underline{u}^T(t)R(t)\underline{u}(t) \} dt, \qquad (3)$$

where $Q(t)$ is a symmetric positive semi-definite real valued $n \times n$ matrix of continuous functions of t defined on $t_0 \leq t \leq t_1$, $R(t)$ is a symmetric positive definite real valued $n \times n$ matrix of continuous functions of t defined on $t_0 \leq t \leq t_1$, P is a symmetric positive semi-definite real valued $n \times n$ matrix of constants and superscript T denotes the transpose of a matrix or vector. The problem is to find $\underline{u}(t)$ on $t_0 \leq t \leq t_1$ such that

a) $\underline{x}(t)$ is continuous on $t_0-\tau \leq t \leq t_1$ with (1) satisfied,

b) $\underline{u}(t)$ is continuous on $t_0 \leq t \leq t_1$,

c) $\underline{x}(t) = \underline{g}(t)$ on $t_0-\tau \leq t \leq t_0$,

and

d) J is an absolute minimum.

A control function $\underline{u}(t)$ which satisfies these conditions is called an optimal control. The corresponding value of J is called the optimal performance index.

CONDITIONS FOR THE OPTIMAL CONTROL

The necessary conditions for optimality are given below.

$$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{x}(t-\tau) + C(t)\underline{u}(t), \quad t_0 \leq t \leq t_1, \tag{3a}$$

$$\underline{x}(t) = \underline{g}(t), \quad t_0-\tau \leq t \leq t_0, \tag{4}$$

$$\underline{\lambda}(t_1) = -P\underline{x}(t_1), \tag{5}$$

$$\dot{\underline{\lambda}}(t) = -A^T(t)\underline{\lambda}(t) + Q(t)\underline{x}(t), \quad t_1-\tau \leq t \leq t_1, \tag{6}$$

$$\dot{\underline{\lambda}}(t) = -A^T(t)\underline{\lambda}(t) - B^T(t+\tau)\underline{\lambda}(t+\tau) + Q(t)\underline{x}(t), \quad t_0 < t < t_1-\tau, \tag{7}$$

$$\underline{u}(t) = R^{-1}(t)C^T(t)\underline{\lambda}(t), \quad t_0 \leq t \leq t_1, \tag{8}$$

$$\underline{\lambda}(t) \text{ is continuous on } t_0 \leq t \leq t_1 . \tag{9}$$

The conditions are obtainable from Pontryagin's maximum principle [4], and have been derived using the calculus of variations in [5], where these conditions also have been shown to be sufficient.

DERIVATION OF THE ALGORITHM

In this section a computational algorithm is derived which results in a numerical solution to the optimization problem. By using functional analysis, Aoki [6] has derived a similar algorithm for the case of systems described by ordinary non-time varying differential equations. Here the algorithm for the solution to the optimal regulator problem with differential-difference equation constraints is derived directly from the performance index.

Consider an arbitrary control $\underline{u}_n(t)$ for $t_0 \leq t \leq t_1$. Using (1) and (2), the corresponding $\underline{x}_n(t)$ can be computed for increasing t from

$$\dot{\underline{x}}_n(t) = A(t)\underline{x}_n(t) + B(t)\underline{x}_n(t-\tau) + C(t)\underline{u}_n(t), \qquad t_0 \le t \le t_1, \tag{10}$$

and

$$\underline{x}_n(t) = \underline{g}(t), \quad t_0-\tau \le t \le t_0. \tag{11}$$

Now using (5), (6), and (7), $\underline{\lambda}_n(t)$ can be computed for decreasing t from

$$\underline{\lambda}_n(t_1) = -P\underline{x}_n(t_1), \tag{12}$$

$$\dot{\underline{\lambda}}_n(t) = -A^T(t)\underline{\lambda}_n(t) + Q(t)\underline{x}_n(t), \quad t_1-\tau \le t \le t_1, \tag{13}$$

and

$$\dot{\underline{\lambda}}_n(t) = -A^T(t)\underline{\lambda}_n(t) - B^T(t+\tau)\underline{\lambda}_n(t+\tau) + Q(t)\underline{x}_n(t), \qquad t_0 \le t \le t_1-\tau, \tag{14}$$

with $\underline{\lambda}_n(t)$ continuous on $t_0 \le t \le t_1$. From the solution to (10) and (11) and the arbitrary $\underline{u}_n(t)$ a performance index J_n can be calculated from (3). A method of choosing $\underline{u}_{n+1}(t)$, based on previously calculated information, is now sought which will result in a smaller performance index.

Define

$$\delta\underline{x}_n(t) = \underline{x}_{n+1}(t) - \underline{x}_n(t) \tag{15}$$

and

$$\delta\underline{u}_n(t) = \underline{u}_{n+1}(t) - \underline{u}_n(t). \tag{16}$$

Consider the augmented performance index

$$\begin{aligned} J_n = \frac{1}{2}\underline{x}_n^T(t_1)P\underline{x}_n(t_1) + \int_{t_0}^{t_1} \{&\frac{1}{2}\underline{x}_n^T(t)Q(t)\underline{x}_n(t) \\ &+ \frac{1}{2}\underline{u}_n^T(t)R(t)\underline{u}(t) + \underline{\lambda}_n^T(t)[\dot{\underline{x}}_n(t) - A(t)\underline{x}_n(t) \\ &- B(t)\underline{x}_n(t-\tau) - C(t)\underline{u}_n(t)]\}dt \end{aligned} \tag{17}$$

with J_{n+1} defined in a similar manner. After integrating (17) by parts and using (10), (12), (13), and (14), the following expression for J_{n+1} will result

$$\begin{aligned} J_{n+1} = J_n &+ \int_{t_0}^{t_1} \{[\underline{u}_n^T(t)R(t) - \underline{\lambda}_n^T(t)C(t)]\delta\underline{u}_n(t)\}dt \\ &+ \frac{1}{2}\delta\underline{x}_n^T(t_1)P\delta\underline{x}_n(t_1) + \int_{t_0}^{t_1} \{\frac{1}{2}\delta\underline{x}_n^T(t)Q(t)\delta\underline{x}_n(t) \end{aligned}$$

$$+ \frac{1}{2} \delta \underline{u}_n^T(t) R(t) \delta \underline{u}_n(t) \} dt; \tag{18}$$

if, in addition, $\delta \underline{x}_n(t)$ satisfies

$$\delta \dot{\underline{x}}_n(t) = A(t) \delta \underline{x}_n(t) + B(t) \delta \underline{x}_n(t-\tau) + C(t) \delta \underline{u}_n(t),$$

$$t_0 \leq t \leq t_1, \tag{19}$$

and

$$\delta \underline{x}_n(t) = \underline{0}, \quad t_0 - \tau \leq t \leq t_0. \tag{20}$$

The $\delta \underline{u}_n(t)$ which generates the (n+1)st control using the (n)th control is selected to have the form

$$\delta \underline{u}_n(t) = - \epsilon_n R^{-1}(t)[R(t)\underline{u}_n(t) - C^T(t)\underline{\lambda}_n(t)], \quad t_0 \leq t \leq t_1, \tag{21}$$

where $\epsilon_n > 0$ is an unspecified constant. For notational simplicity define

$$\underline{m}_n(t) = R(t)\underline{u}_n(t) - C^T(t)\underline{\lambda}_n(t). \tag{22}$$

Substituting (21) into (19), and letting

$$\delta \underline{x}_n(t) = \epsilon_n \underline{h}_n(t), \tag{23}$$

the result is

$$\epsilon_n \dot{\underline{h}}_n(t) = \epsilon_n A(t)\underline{h}_n(t) + \epsilon_n B(t)\underline{h}_n(t-\tau) - \epsilon_n C(t)R^{-1}(t)\underline{m}_n(t), \tag{24}$$

and from (20)

$$\epsilon_n \underline{h}_n(t) = \underline{0}, \quad t_0 - \tau \leq t \leq t_0.$$

Since ϵ_n is taken to be non-zero, $\underline{h}_n(t)$ must satisfy

$$\dot{\underline{h}}_n(t) = A(t)\underline{h}_n(t) + B(t)\underline{h}_n(t-\tau) - C(t)R^{-1}(t)\underline{m}_n(t) \tag{25}$$

$$\text{for } t_0 \leq t \leq t_1;$$

and

$$\underline{h}_n(t) = \underline{0}, \quad t_0 - \tau \leq t \leq t_0. \tag{26}$$

Therefore, $\underline{h}_n(t)$ can be calculated from the initial choice of $\underline{u}_n(t)$. Substituting (21) and (23) into (18), the result is

$$J_{n+1} = J_n - \epsilon_n \int_{t_0}^{t_1} \underline{m}_n^T(t) R^{-1}(t) \underline{m}_n(t)\,dt + \frac{1}{2} \epsilon_n^2 \underline{h}_n^T(t_1) P \underline{h}_n(t_1)$$

$$+ \frac{1}{2} \epsilon_n^2 \left[\int_{t_0}^{t_1} \underline{h}_n^T(t) Q(t) \underline{h}_n(t)\,dt + \int_{t_0}^{t_1} \underline{m}_n^T(t) R^{-1}(t) \underline{m}_n(t)\,dt \right]. \tag{27}$$

Define

$$d = \int_{t_0}^{t_1} \underline{m}_n^T(t) R^{-1}(t) \underline{m}_n(t)\,dt, \tag{28}$$

$$r = \int_{t_0}^{t_1} \underline{h}_n^T(t) Q(t) \underline{h}_n(t)\,dt, \tag{29}$$

$$s = \underline{h}_n^T(t_1) P \underline{h}_n(t_1). \tag{30}$$

Then (27) becomes

$$J_{n+1} = J_n - \epsilon_n d + \frac{1}{2} \epsilon_n^2 (s+r+d). \tag{31}$$

Now $s \geq 0$ since P is positive semi-definite, $r \geq 0$ since Q is positive semi-definite, and $d > 0$, unless $\underline{m}_n(t) \equiv 0$ on $t_0 \leq t \leq t_1$, since R^{-1} is positive definite. Note that $\underline{m}_n(t) \equiv 0$ implies that Equation (8) is satisfied; that is, $\underline{u}_n(t)$ is an optimal control. From (31), it is seen that any choice of ϵ_n in the range $0 < \epsilon_n < 2d/s+d+r$ will result in

$$J_{n+1} < J_n . \tag{32}$$

Since J_n, n=1,2,3,... are bounded below by zero the sequence possesses a limit point. More information can be obtained from (31). The value of ϵ_n which results in the greatest decrease of the performance index is

$$\epsilon_n = \frac{d}{s+r+d}, \tag{33}$$

and the resulting change for this value of ϵ_n is

$$\Delta J_n = J_{n+1} - J_n = -\frac{1}{2} \frac{d^2}{s+r+d} . \tag{34}$$

Equation (34) is useful for comparing theoretically predicted changes in J with the actual changes which result when this algorithm is used on the computer.

SUMMARY OF THE ALGORITHM

A summary of the steps involved in the computer solution to the problem is given below.

1. Choose $\underline{u}_1(t)$ for $t_0 \leq t \leq t_1$ and store. Most conveniently, $\underline{u}_1(t)$ is taken to be identically zero.
2. Calculate $\underline{x}_1(t)$ for increasing t from

   $$\dot{\underline{x}}_1(t) = A(t)\underline{x}_1(t) + B(t)\underline{x}_1(t-\tau) + C(t)\underline{u}_1(t) \text{ and}$$

   $x_1(t) = \underline{g}(t)$, $t_0-\tau \leq t \leq t_0$. At the same time calculate J_1 and store $\underline{x}_1(t)$.
3. Calculate $\underline{\lambda}_1(t)$ for decreasing t from

   $$\underline{\lambda}_1(t_1) = -P\underline{x}_1(t_1),$$

   $$\dot{\underline{\lambda}}_1(t) = -A^T(t)\underline{\lambda}_1(t) + Q(t)\underline{x}_1(t), \quad t_1-\tau \leq t \leq t_1, \text{ and}$$

   $$\dot{\underline{\lambda}}_1(t) = -A^T(t)\underline{\lambda}_1(t) - B^T(t+\tau)\underline{\lambda}_1(t+\tau) + Q\underline{x}_1(t), \quad t_0 \leq t \leq t_1-\tau.$$

   At the same time calculate $\underline{m}_1(t) = R(t)\underline{u}_1(t) - C^T(t)\underline{\lambda}_1(t)$ and store.
4. Calculate for increasing t

   $$\dot{\underline{h}}_1(t) = A(t)\underline{h}_1(t) + B(t)\underline{h}_1(t-\tau) - C(t)R^{-1}\underline{m}_1(t) \text{ and}$$

   $\underline{h}_1(t) = \underline{0}$ for $t_0-\tau \leq t \leq t_0$. At the same time, evaluate

   $$d = \int_{t_0}^{t_1} \underline{m}_1^T(t)R^{-1}(t)\underline{m}_1(t)\,dt,$$

   $$r = \int_{t_0}^{t_1} \underline{h}_1^T(t)Q(t)\underline{h}_1(t)\,dt,$$

   and

   $$s = \underline{h}_1^T(t_1)P\underline{h}_1(t_1).$$
5. Calculate $\epsilon_1^{op} = \dfrac{d}{s+r+d}$ and $J_2 - J_1 = -\dfrac{1}{2}\dfrac{d^2}{s+r+d}$.
6. Calculate $\underline{u}_2(t) = \underline{u}_1(t) + \delta\underline{u}_1(t) = \underline{u}_1(t) - \epsilon_1^{op} R^{-1}\underline{m}_1(t)$

   and store $\underline{u}_2(t)$ in the location of $\underline{u}_1(t)$.
7. Continue through the sequence using $\underline{u}_2(t)$ as the starting control.
8. Stop when d becomes sufficiently small.

With this method only one n-vector, $\underline{x}(t)$, and two r-vectors, $\underline{u}(t)$ and $\underline{m}(t)$ must be stored. A successful iteration, that is, one which results in a smaller performance index, is guaranteed. Therefore, the storage of two successive controls, which is necessary when methods based on only the first variation [7] are used, is not needed when this algorithm is employed.

A program has been written for the CDC 1604 computer which follows this algorithm. The method used for solving the differential-difference equations on the computer is outlined by Bellman [8]. With this method a standard differential equation sub-routine can be used; and the storage is minimized. The computer program has been used to obtain solutions for several examples which are contained in [5].

REFERENCES

1. Krasovskii, N. N., "On the Analytical Design of an Optimal Controller in Systems with Time Lags," Prikl. Matem. i. Mekh., Vol. 26, No. 1, 1962.
2. Krasovskii, N. N., "Optimal Processes in Systems with Time Lags," Proceedings of IFAC, Paper No. 52911, 1963.
3. Kramer, J.D.R., "On Control of Linear Systems with Time Lags," Information and Control, Vol. 3, pp. 299-326, 1960.
4. Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mischenko, E. F., The Mathematical Theory of Optimal Processes, Interscience Publishers, pp. 213-226, 1962.
5. Mueller, T. E., "Optimal Control of Linear Systems with Time Lag," Technical Report R-254, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, June 1965.
6. Aoki, M., "On a Successive Approximation Technique in Solving Some Control System Optimization Problems," Journal of Mathematical Analysis and Applications, Vol. 5, pp. 418-434, 1962.
7. Merriam, C. W., Optimization Theory and the Design of Feedback Control Systems, McGraw-Hill Book Co., 1964.
8. Bellman, R., "On the Computational Solution of Differential-Difference Equations," Journal of Mathematical Analysis and Applications, Vol. 2, pp. 108-110, 1961.

CONTROLLABILITY OF LINEAR SYSTEMS WITH TIME LAG

G.S.TAHIM
Assistant Professor, Electrical Engineering Dept.
University of Kansas, Lawrence, Kansas

The problem of optimal control can be stated in the following way - Can the system be transferred from one given state to the desired final state in such a way that the behaviour of the system satisfies certain cost restrictions. The first thing to be considered is that whether the desired transference of the system from one state to the other is possible in a finite length of time by application of the control effort. The concept of controllability introduced by Kalman [1] can be used to answer this important question. The condition for controllability of linear system governed by ordinary differential equations are well known. However these criterion when the linear system we are considering has a memory requirement are not known. Krasovki [2] solved an optimization problem for systems involving time lags. These systems are governed by a system of linear differential difference equations. Krasovski assumed the existence of the solution of the optimization problem. The controllability conditions metioned below assure the existence.

Consider the system of linear differential-difference equations:

$$x(t) = A\,x(t) + B\,x(t-\rho) + C\,u(t)\ , \quad t > 0 \tag{1}$$

The initial function is given as

$$x(t) = g(t) \qquad -\rho \leqq t \leqq 0 \tag{2}$$

$\rho > 0$ is the time lag. A,B, and C are constant matrices of order $n \times n$, $n \times n$, and $n \times r$ respectively. x(t) an n - vector represents the state of the system. u(t), an r-vector, represents the control. In general, in the rest of the paper, capital letters represent matrices and vector spaces, lower case letters represent vectors and greek letters stand for scalars. Prime represents a transpose. $C^n\,[a,b]$ represents the class of real functions defined on $[a,b]$ such that its derivatives upto and including nth order are continuous. Dot represents time derivative and r(A) denotes the rank of matrix A.

$$g(t) \in C^0\,[-\rho, 0] \tag{3}$$

$$g_c = \text{subclass of } g(t) \text{ such that } g(t) = \alpha\ , \ -\rho \leqq t \leqq 0.$$

Obviously g_c's span E_n.

We wish to write the solution of (1) for $t > 0$ such that $x(t) = g(t)$, $-\rho \leqq t \leqq 0$. The existence of a unique solution of (1) with conditions (2) and (3) has been proved by Bellman and Cooke [3]. To get the solution in explicit form we rewrite (1) as below:

$$x(t) = A\,x(t) + f(t) \tag{4}$$

where f(t) is a forcing function defined by:

$$f(t) = B\, g(t-\rho) + C\, u_1(t), \quad o \leqq t \leqq \rho \tag{5}$$

and $u_1(t)$ = control vector defined over $o \leqq t \leqq \rho$

The solution of (5) is given by [4]

$$x(t) = \Phi(t)\, x(o) + \int_o^t \Phi(t-\tau)\, f(\tau) d\tau, \quad o \leqq t \leqq \rho \tag{6}$$

where $\Phi(t) = e^{At}$ is the transition matrix of the equation $\dot{x}(t) = A\, x(t)$
Using (5) and (6) we obtain

$$x(t) = \Phi(t) \left[x(o) + \int_o^t \Phi(-\tau)\, B\, g(\tau-\rho)\, d\tau + \int_o^t \Phi(-\tau)\, C\, u_1(\tau) d\tau \right], \quad o \leqq t \leqq \rho \tag{7}$$

Define

$$\Phi(-t)\, B\, g(t-\rho) = f_g(t) - \tag{8}$$

and

$$\Phi(-t)\, C\, u_1(t) = f_1(t) \tag{9}$$

(7) can now be written as

$$x(t) = \Phi(t) \left[x(o) + \int_o^t f_g(\tau) d\tau + \int_o^t f_1(\tau) d\tau \right], o \leqq t \leqq \rho \tag{10}$$

Let $u_2(t)$ = control vector defined over $\rho \leqq t \leqq 2\rho$. We can now treat (10) as the initial function defined over $0 \leqq t \leqq \rho$ and $u_2(t)$ as the control vector over $[\rho, 2\rho]$; this enables us to write the solution of (1) over $[\rho, 2\rho]$. Iteratively we can extend the solution over the next lag interval.

DEFINITION: The system (1) is said to be completely controllable if for any $g(t) \in C^o([-\rho, o], R^n)$ there exists a T, $T \geqq \epsilon > o$ and a piecewise continuous control vector u(t) defined for $t > 0$ such that $x(t) = 0$ over $T \leqq t \leqq T + \rho$.

We now give the main result:

THEOREM: The system (1) is completely controllable if and only if the following conditions are satisfied:

(a) $B = C\, V^T$ where V is an arbitrary $n \times r$ matrix of constants

(b) rank $[C, AC, A^2C, \ldots\ldots, A^{n-1}C] = n$.

PROOF: (sufficiency) IN (7) substitute $B = CV^T$ and rewrite

$$x(t) = \Phi(t)\left[x(o) + \int_0^t \Phi(-\tau)\, CV^T g(\tau-\rho)d\tau + \int_0^t \Phi(-\tau)\, Cu_1(\tau)d\tau\right], \quad o \leqq t \leqq \rho \qquad (11)$$

Let $u_1(t) = u_{1A}(t) + u_{1B}(t)$ (12)

Using (12) in (11) and rearranging

$$x(t) = \Phi(t)\left[x(o) + \int_0^t \Phi(-\tau)\, C\left\{V^T g(\tau-\rho) + u_{1B}(\tau)\right\}d\tau + \int_0^t \Phi(-\tau)\, Cu_{1A}(\tau)\, d\tau\right] \qquad (13)$$

Choose $u_{1B}(t) = -V^T g(t-\rho), \quad o \leqq t \leqq \rho$ (14)

From (14) and (13)

$$x(t) = \Phi(t)\left[x(o) + \int_0^t \Phi(-\tau)\, C u_{1A}(\tau)d\tau\right], \quad o \leqq t \leqq \rho \qquad (15)$$

Define $\dot{x}(t) = A\, x(t) + C\, u(t)$ (16)

From [5] it is known that if condition (b) of Theorem 1 is satisfied in (15), there exists a $T \geqq \epsilon > o$ and a piecewise continuous control such that $x(T) = 0$ for all $x(0)$ and hence for all $g(t)$. To keep $x(t) = 0$ over the interval $[T, T+\rho]$ we select the following control vectors.

$$u_{1A}(t) = o \qquad T \leqq t \leqq T+\rho$$

$$u_{1B}(t) = \begin{cases} -V^T g(t-\rho) & T \leqq t \leqq T+\rho \\ -V^T \varphi_u(t-\rho\,;\, x(o),\, o) & \rho \leqq t \leqq T+\rho \end{cases}$$

where $\varphi_u(t;\, x(0), 0)$ = trajectory of (16) at t, starting from x(0) at time t = 0 and control $u_{1A}(t)$ as defined in [5] mentioned above.

(Necessity) We make use of the following theorem from the theory of matrices [7] to see the method of attack.
Theorem: Two matrices whose elements are explicitly given numbers can be transformed each into the other by elementary transformations if and only if they have the same rank and the same order.

Thus to establish the factorization of matrix B we have to show the following:

(A) $r(B) = r(C)$
(B) If $r(B) = r(C)$ but B can not be factored as $B{*}CV^T$ then there exist uncontrollable states.
(C) Lastly we establish that if condition (b) is not satisfied then there exist no T 0 such that $x(T) = 0$ for all $g(t)$.

(A). Let $r(B) = r_2$, $r(C) = r_1$. Suppose $r_1 \neq r_2$. e_j is an n-vector with zeros everywhere except in j th row. Apply a nonsingular transformation T to (1) such that the image space of B has a basis $e_1, e_2, \ldots\ldots, e_{r2}, 0, 0, \ldots\ldots 0$. Let $g_{cj} \in g_c(t)$, $j = 1, 2, \ldots\ldots n$. α_{ji} are constants. Now,

$$\int_0^t \Phi(-\tau)\, B\, g_{cj}\,(\tau-\rho)\, d\tau = \int_0^t \Phi(-\tau) \left[\sum_{i=1}^{r_2} \alpha_{ji}\, e_i \right] d\rho$$

= a vector which has zeros in the rows $r_2+1, r_2+2, \ldots\ldots, n$. (17)

Also,

$$\int_0^t \Phi(-\tau)\, C\, u_1(\tau)\, d\tau = \int_0^t \Phi(-\tau)\, K_1^{-1} (K_1 C K_2) K_2^{-1}\, u_1(\tau)\, d\tau$$

$$= \int_0^t \Phi(-\tau)\, K_1^{-1} \left[e_1, \ldots, e_{r1}, \ldots 0, 0 \right].$$

$$K_2^{-1}\, u_1(\tau)\, d\tau \; .$$

= a vector with zeros in rows $r+1, \ldots$ $\ldots, n$ for any control, if we choose $\Phi(-\tau)\, K_1^{-1}$ with zero entries for

$i = r_1+1, \ldots\ldots, n$ $j = 1, 2, \ldots, r1$.

(18)

In (18) the existence of nonsingular K_1 and K_2 is assured since $r(C) = r_1$. From (10) we know that (17) represents the effect of initial function on solution while (18) represents the effect of control on the solution. Comparing (17) and (18) it is seen that not all initial functions can be controlled unless $r_1 = r_2$.

(B). For simplicity we assume $r(C) = r_1 = r$ i.e., C has maximal rank. The proof can be extended to the case where $r_1 \quad r$. Apply a nonsingular transformation T to (10) so that C appears as

$$C = \begin{bmatrix} 0 \\ \hline I_r \end{bmatrix} \tag{19}$$

C, 0 and I_r are matrices of order $n \times r$, $(n-r) \times r$ and $r \times r$ respectively. $I_n =$ identity matrix of order n. Partition $V^T = \left[V_1 \mid V_2 \right]^T$ where V^T is an $r \times n$ matrix of constants, V_1^T is an $r \times (n-r)$ matrix and V_2^T an $r \times r$ matrix. Now

$$CV^T = \begin{bmatrix} 0_1 & 0_2 \\ V_1^T & V_2^T \end{bmatrix} \tag{20}$$

where 0_1 and 0_2 are matrices of zeros of order $n-r \times n-r$ and $n-r \times r$ respectively. Assume $B \neq CV^T$. Then for some i,j; $b_{ij} \neq 0$; $i = 1,2,.n-r$, $j = 1,2,\ldots,n$. Let $g_1 \in g_c$ be such that $g_1 = e_j$ then $\left[Bg_1 \right]_{i,j} = b_{ij}$

$$\int_0^t \Phi(-\tau)\, B\, g_1\, d\tau \quad = \text{an n-vector which for any A has a nonzero entry in the } i^{th} \text{ row.} \tag{21}$$

Choose an A such that $\Phi(-\rho)$ has zero entries in the last r columns of the i^{th} row. Then for any $u_1(\tau)$,

$$\int_0^t \Phi(-\tau)\, C\, u_1(\tau) d\tau = \text{an n-vector with zero entry in the } i^{th} \text{ row.} \tag{22}$$

Comparing (21) and (22) we see that for some A and initial function e_j the control has no effect on some states if $b_{ij} \neq 0$. Therefore $b_{ij} = 0$ and $B = C\, V^T$.

(C). $r\left[C, AC, \ldots\ldots. A^{n-1} C\right] \neq n$ implies there is no $t > 0$ such that over $[0,t]$ rows of $\Phi(-t)C$ are linearly independent .

Thus there exists a vector such that

$$y^T \Phi(-t)\, C = o \quad \text{for all } t > o \tag{23}$$

Premultiply (7) by $y^T \Phi(-t)$

$$y^T \Phi(-t)\, x(t) = y^T \Big[x(o) + \int_0^t \Phi(-\tau)\, B\, g(\tau - \rho)\, d\tau$$

$$+ \int_0^t \Phi(-\tau)\, C\, u_1(\tau)\, d\tau \Big], o \leq t \leq \rho \tag{24}$$

Using (23) in (24) and letting g(o) = y, we get

$$y^T \Phi(-t)\, x(t) = y^T \Big[y + \int_0^t \Phi(-\tau)\, B\, g(\tau - \rho)\, d\tau \Big] \tag{25}$$

From (25) we see that there is no T in the interval $[o, \rho]$ such that x(T) = 0. It is easy to see that if instead of (7) we put the expression for x(t) over $[\rho, 2\rho]$ we can assert there is no $T \in [\rho, 2\rho]$ such that x(T) = 0; iterartively we can show that there is no T in any subsequent interval such that x(T) = 0.

But the system is completely controllable. The contradiction established that $r \big[C, AC, \ldots\ldots A^{n-1} C \big] = n$. This completes the proof of Theorem 1.

Thus we have established the necessary and sufficient conditions for complete controllability of a system governed by linear differential-difference equations with constant coefficients. The conditions requrie that the system with the matrix weighing the delayed states set to zero, should be completely controllable and the matrix weighing the delayed states should be factorable as a product of the matrix weighting the control and an arbitrary matrix of constants. These conditions assure the existence of a solution to the optimization problem.

BIBLIOGRAPHY

1. Kalman, R.E., On the General Theory of Control Systems, Proceedings of 1st International Congress on Automatic Control, Moscow, 1960, Butterworths Scientific Publications, London, Vol; 1, p.481, 1961.

2. Krasovskii, N.N., On the Analytic Construction of an Optimal Control in a system with Time Lags, Prik. Mat. I Mek. Vol, 26, No. 1, p. 50-67, 1962.

3. Bellman, R. and Cooke, K.L., Differential-Difference Equations Academic Press, New York, 1963, p. 171.

4. Coddington, E.A. and Levinson, N., Theory of Ordinary Differential Equations, McGraw - Hill Book Co. Inc., New York, 1955, Chapter 3.

5. Kalman, R.E., Contributions to the Theory of Optimal Control, Bol. Soc. Mat. Mexicana, pp. 102-119, 1960.

6. Weiss, L. and Kalman R.E., Contributions to Linear System Theory RIAS Technical Report 64-9, RIAS, Baltimore, April, 1964.

7. Hohn, Franz, E., Elementary Matrix Algebra, The Macmillan Book Company, New York, 1957, pp. 92.

CONTROLLABILITY OF A CLASS OF NONLINEAR SYSTEMS WITH BOUNDED STATE AND CONTROL SPACES

WILLIAM D. O'NEILL
Department of Systems Engineering
University of Illinois at Chicago Circle

ABSTRACT

This paper is concerned with the controllability of a general class of nonlinear systems describable by an n^th order ordinary differential equation. The state and control spaces of the system variables are assumed to be bounded by inequality constraints. It is shown that if the system equations are defined only within the restricted state and control spaces then easily determinable regions of uncontrollability arise. The states in the uncontrollable regions satisfy a sufficient condition of uncontrollability relative to states that are in the bounded region of the control space but not in the uncontrollable region. An example is given to illustrate the theoretical results.

I. INTRODUCTION

The recent years have seen considerable effort expended in advancing the theory of optimal control of dynamical systems using a variety of variational techniques [6], [8], [9], [10], [11]. A common practical drawback shared by these methods is the local nature of the optimal solutions resulting from their employment. Another applicational difficulty present in most of these techniques is the lack of simple means to determine a priori if the system is even controllable.

In order to avoid such difficulties, several Authors have investgated optimization procedures in the presence of bounded control and state spaces but with limited practical results [1], [2], [3] except perhaps those of [4], [5].

It is the purpose of this paper to consider a class of physically significant systems whose state and control spaces are bounded by inequalities. There are transformation techniques available that make it possible to reformulate the bounded problem in the format of an unbounded problem but the actual computational difficulties arising from such transformations usually render solutions possible only in relatively simple situations [7].

It will be shown in the ensuing sections that under practical assumptions about the systems' mathematical description, the state and control space bounds actually help in determining a priori certain regions in the restricted state space that are controllable. If the system begins in or is driven into an uncontrollable region then the results of this paper show that there does not exist an admissible control capable of taking the state to any state exterior to the uncontrollable region.

II. PROBLEM FORMULATION AND ASSUMPTIONS

The class of systems to be considered in this paper is assumed to be representable by the n^th order equation

$$\frac{d^n y}{dt^n} = f\left(y, \frac{dy}{dt}, \ldots, \frac{d^{n-1}y}{dt^{n-1}}; u_1, u_2, \ldots, u_r\right) \tag{1}$$

where y is the system output variable and u_1, $u_2, \ldots, u_r$ the control variables. This equation is easily transformed into the normal form

$$\begin{aligned} x_1 &= y \\ \dot{x}_1 &= x_2 \\ &\vdots \\ \dot{x}_n &= f(x,u) \end{aligned} \tag{2}$$

where $x = \text{col}(x_1, x_2, \ldots, x_n)$ and $u = \text{col}\ (u_1, u_2, \ldots, u_r)$

For the system (2) the following assumptions are made:

1) The components of the state vector x are restricted by the inequalities

$$|x_i| \leq N_i \quad ; \quad i = i, 2, \ldots, n.$$

2) The control vector components are restricted by the inequalities

$$|u_j| \leq M_j \quad ; \quad j = 1, 2, \ldots, r.$$

3) Equations (2) are valid only in the control and state space region described in 1) and 2).

4) For x and u within the restricted region described by 1) and 2), $|f(x,u)|$ is bounded by some number B.

These assumptions, although mathematical in form, are seen to be met in many realistic problem formulations intended to represent a physical system. Assumptions 1), 2) and 4) are intended to eliminate unrealistically large signals in the system while 3) is justified by the observation that system behavior at or beyond the limit of component durability is often difficult if not impossible to describe mathematically. In a specific situation the numbers N_i and M_j arise from physical considerations of the system components e.g. rated motor voltages, mechanical stops, etc.

III. UNCONTROLLABLE REGIONS

In order to proceed with the mathematical description of the uncontrollable regions of (2), it is convenient to define the point set

$$G = \left\{ x \mid x \in \bigcap_{i=i}^{n} \left[x ; |x_i| \leq N_i \right] \right\} \tag{3}$$

Quite obviously then, G is the restricted region in the state space to which trajectories of (2) must be confined.

The uncontrollable regions interior to G are derived using the equations (2) and the system constraints. From (2)

$$x_n \frac{dx_n}{dx_{n-1}} = f(x,u) \tag{4}$$

and by the assumption $|f(x,u)| \leqq B$ so that (4) implies

$$-B \leqq x_n \frac{dx_n}{dx_{n-1}} \leqq B \quad . \tag{5}$$

Now let $\xi = \mathrm{col}(x_1(t_o), x_2(t_o),\ldots,x_n(t_o))$ be any initial state in G. Integration of (5) from ξ to x(t) yields

$$-2B(x_{n-1} - \xi_{n-1}) + \xi_n^2 \leqq x_n^2 \leqq 2B(x_{n-1} - \xi_{n-1}) + \xi_n^2. \tag{6}$$

For $x_n > 0$ a geometrical interpretation of (6) is given in Fig. 1. The analytical interpretation of the same inequality is that for some initial state $\xi \in$ G and any admissible control vector, the projections of all possible trajectories leaving ξ onto the x_n, x_{n-1} plane must remain in the shaded region of Fig. 1 for some time $t_o \leqq t \leqq T_n$. Here T_n represents the first time after t_o that it is possible to have a state outside of the region described by (6). Proceeding in much the same manner as above, the validity of the inequalities

$$-N_i \leqq x_{i-1} \frac{dx_{i-1}}{dx_{i-2}} \leqq N_i \quad ; \quad i = 3,4,\ldots,n \tag{7}$$

is directly verified from (2). Integration of (7) gives the inequalities

$$-2N_i(x_{i-2} - \xi_{i-2}) + \xi_{i-1}^2 \leqq x_{i-1}^2 \leqq 2N_i(x_{i-2} - \xi_{i-2}) + \xi_{i-1}^2 \tag{8}$$

$$i = 3,4,\ldots,n.$$

Comparison of (8) and (6) indicates that (8) results in the same type regions as those of Fig. 1 but now in the x_{i-1}, x_{i-2} planes. For n = 3 the situation is illustrated in Fig. 2.

To arrive at the uncontrollable regions by the above, consider first inequalities (6) for initial states such that

$\xi_n = \pm N_n$, $\xi_{n-1} = \pm\left(N_{n-1} - \frac{N_n^2}{2B}\right)$. For these initial states the shaded regions of Fig. 3 result. As is easily verified by direct application of (6), trajectories starting at any state in the shaded regions result in projections that are still in the shaded regions. The results of the Appendix show that trajectories on the planes $x_{n-1} = \pm N_{n-1}$ are inconsistent with the physical assumptions

(Section II) unless $x_n = 0$. This same result is true for the planes $x_i = \pm N_i$; $i = 3, 4, \ldots, n-1$ so that use of (8) for $i = 3, 4, \ldots, n$ yields regions similar to those of Fig. 3 but in the x_{i-1}, x_{i-2}; $i = 3, 4, \ldots, n$ planes. Thus the projected trajectories can never leave the shaded regions without leaving G. Consequently, the uncontrollable region in the n space of x can be concisely expressed by unions and intersections of the following sets. Define

$$\begin{aligned}
S_n &= \{x \mid x_n \geqq 0, \; x_n > [-2B \; (x_{n-1} - N_{n-1})]^{\frac{1}{2}}\} \\
S_{n-1} &= \{x \mid x_{n-1} \geqq 0, \; x_{n-1} > [-2N_n \; (x_{n-2} - N_{n-2})]^{\frac{1}{2}}\} \\
&\;\;\vdots \\
S_2 &= \{x \mid x_2 \geqq 0, \; x_2 > [-2N_2 \; (x_1 - N_1)]^{\frac{1}{2}}\} \\
S'_n &= \{x \mid x_n \leqq 0, \; x_n < -[-2B(x_{n-1} - N_{n-1})]^{\frac{1}{2}}\} \\
S'_{n-1} &= \{x \mid x_{n-1} \leqq 0, \; x_{n-1} < - [-2N_n \; (x_{n-2} - N_{n-2})]^{\frac{1}{2}}\} \\
&\;\;\vdots \\
S'_2 &= \{x \mid x_2 \leqq 0, \; x_2 < - [-2N_2 \; (x_1 - N_1)]^{\frac{1}{2}}\}
\end{aligned}$$

Use of these definitions shows the uncontrollable region, I, is given by

$$I = G \bigcap_{i=2}^{n} (S_i \cup S'_i).$$

It is interesting to note that if assumption 4 (Section II) is violated e.g., by allowing unbounded control signals, then I is empty.

To see this, note that the curves $x_n = (\pm 2B(x_{n-1} - N_{n-1}))^{\frac{1}{2}}$ (Fig. 3) intersect $x_n = \pm N_n$ at $x_{n-1} = \pm (N_{n-1} - \frac{N_n^2}{2B})$. Therefore as $B \rightarrow \infty$, $N_{n-1} - \frac{N_n^2}{2B} \rightarrow N_{n-1}$ and $S_n \cup S'_n$ is empty hence I is empty.

Thus if $x \in I$ then x is uncontrollable but if a state in G is uncontrollable this does not necessarily mean that $x \in I$ as numerous examples in [12] show. The following example illustrates a case where $x \in I$ is both a necessary and sufficient condition for x to be uncontrollable.

Example.

Consider the second order system

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = u$$

With control and state constraints

$$|x_1| \leqq 3, \quad |x_2| \leqq 2, \quad |u| \leqq \tfrac{1}{2}.$$

Let it be required to reach the origin in minimum time from some initial state $(x_1(0), x_2(0)) \in G$. For this same system with only bounded control the time optimal trajectories and control are well known (6).

To determine the uncontrollable regions inside of $G = \{x_1, x_2 \mid |x_1| \leqq 3, |x_2| \leqq 2\}$, inequality (6) is applied with $B = \frac{1}{2}$ for initial states $\xi_1 = \pm 1$, $\xi_2 = \pm 2$. The resultant I region is shown shaded in Fig. 4. The time optimal trajectories are also shown to illustrate that the I region for this example represents states that are necessarily and sufficiently uncontrollable.

IV CONCLUSION

In this paper it is shown that for systems that are described by equations of the form (2), constraints on the control and state spaces give rise to easily determinable uncontrollable regions in the restricted state space. Under practical assumptions, the uncontrollable regions are shown to have the same geometric shape for any system of the form (2) since the analytic equations of the boundaries depend only on the state and control space bounds and the bound on the n^{th} time derivative of the output.

It should be noted that uncontrollable regions for systems described by (2) but with different state space bounds can also be obtained using the technique outlined in this paper so long as the region in the state space is bounded.

APPENDIX

It remains to be shown that trajectories of (2) can only exist on restricted sets of the state space boundaries and still be consistent with the assumptions (Section II) and constraints placed on the system.

To show the above result, consider what the equations (2) require for a trajectory to be on the plane $x_i = N_i$, $i = 1, 2, \ldots, n-1$. Obviously $\dot{x} = 0$ and $x_{i+1} = 0$. An additional requirement is derived from (2) by considering what $f(x, u)$ must be if a trajectory is to exist on the plane $a_i (x_i - N_i) + a_n x_n = 0$. Since the gradient of this plane must be orthogonal to the velocity vector of (2), it must be the case that

$$f(x, u) = \frac{-a_i x_{i+1}}{a_n} \quad . \tag{A1}$$

Now $a_i = 1$ and $a_n = 0$ when $x_i = N_i$ so that for $x_{i+1} \neq 0$, (A1) requires the assumption of a bounded $f(x,u)$ to be violated. Hence trajectories on the planes $x_i = N_i$, $i = 1, 2, \ldots, n-1$ are restricted to the lines $x_{i+1} = 0$. For the case $x_n = N_n$, trajectories are confined to the set x, u such that $f(x,u) = 0$. As can be seen from the definitions (9) the boundary trajectories allowable by the above analysis do not afford possible "escape routes" from the uncontrollable regions derived in the text.

REFERENCES

1. R. V. Gamkrelidze, "Optimal Processes with Restricted Phase Coordinates," Bull. Acad. Sci. URSS. Ser. Mat., 24, 1960.

2. S. S. L. Chang, "Optimal Control in Bounded Phase Space," Automatica, I, pp. 55-67, 1962.

3. S. S. L. Chang, "A Modified Maximum Principle for Optimum Control of a System with Bounded Phase Coordinates," AFOSR Report No. 1238, 1961.

4. R. E. Bellman and S. E. Dreyfus, Applied Dynamic Programming, Princeton University Press, Princeton N. J., 1962.

5. Y. C. Ho and P. B. Brentani, "On Computing Optimal Control with Inequality Constraints," Journal on Control, SIAM Publishers, Philadelphia, 1963.

6. L. S. Pontryagin, et al, The Mathematical Theory of Optimal Processes, Interscience Publishers, N.Y., 1962.

7. Optimization Techniques, Edited by G. Leitmann, Academic Press, N.Y. 1962.

8. J. P. LaSalle, "Time Optimal Control Systems," Contributions to the Theory of Non-Linear Oscillations, Ann. Math. Stud., 5, Princeton University, Princeton N. J., 1960

9. L. D. Berkovitz, "Variational Methods in Problems of Control and Programming," Journal of Mathematical Analysis and Applications, 3, I, Aug. 1961, pp. 145-169.

10. S. E. Dreyfus, "Dynamic Programming and the Calculus of Variations," Journal of Mathematical Analysis and Applications, I, 2, pp. 228-239, 1960.

11. R. E. Kalman, Mathematical Optimization Techniques, Chap. 16, edited by R. E. Bellman, RAND Corp. Berkeley, 1963.

12. W. D. O'Neill, "Controllability and Time Optimality of Systems with Linear Control and Bounded State Space," University of Notre Dame Department of Electrical Engineering Report No. EE. 655, Notre Dame, Indiana, 1965.

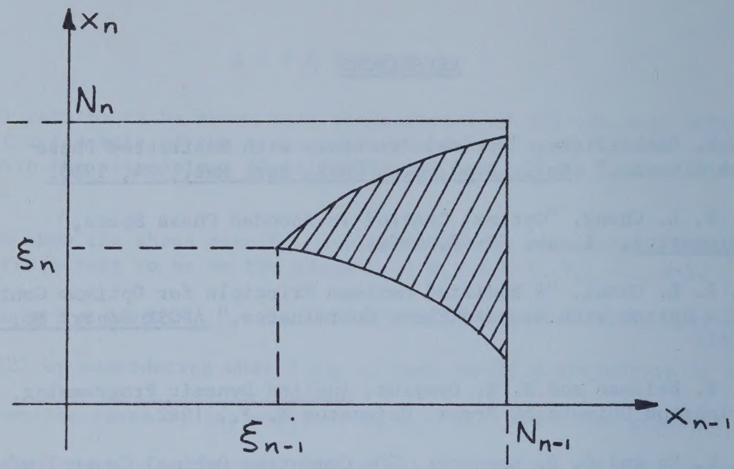

x_n
N_n
ξ_n
ξ_{n-1}
N_{n-1}
x_{n-1}

Figure 1
Geometrical Interpretation of Inequality (6).

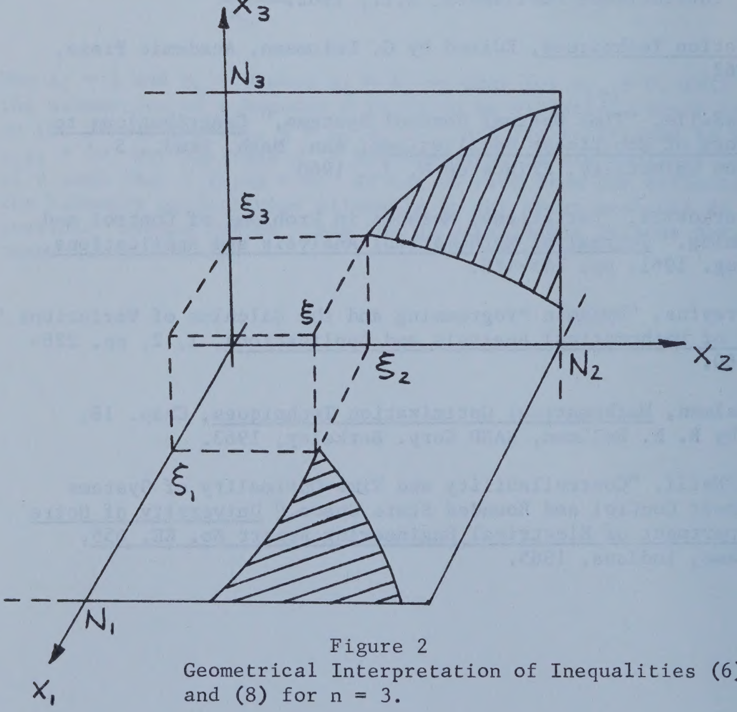

x_3
N_3
ξ_3
ξ
ξ_2
N_2
x_2
ξ_1
N_1
x_1

Figure 2
Geometrical Interpretation of Inequalities (6)
and (8) for n = 3.

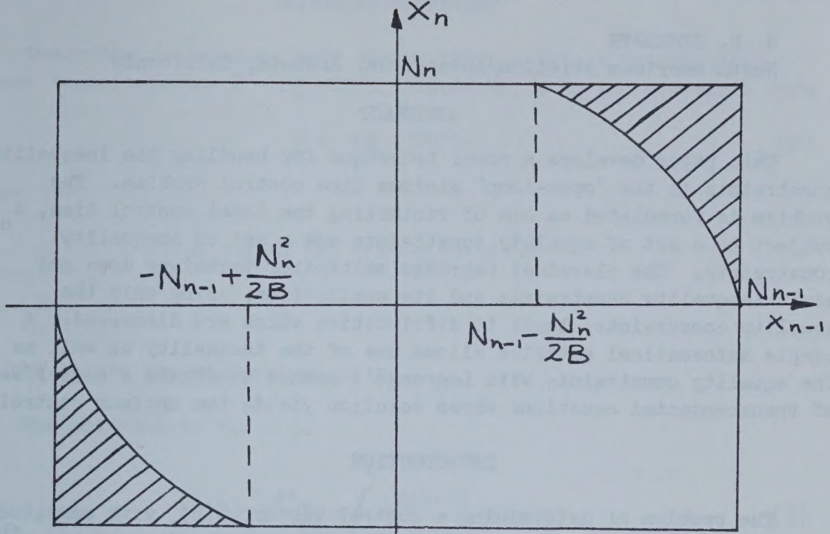

X_n
N_n
$-N_{n-1}+\frac{N_n^2}{2B}$
N_{n-1}
$N_{n-1}-\frac{N_n^2}{2B}$
x_{n-1}

Figure 3
Uncontrollable Region in the x_n,x_{n-1} plane.

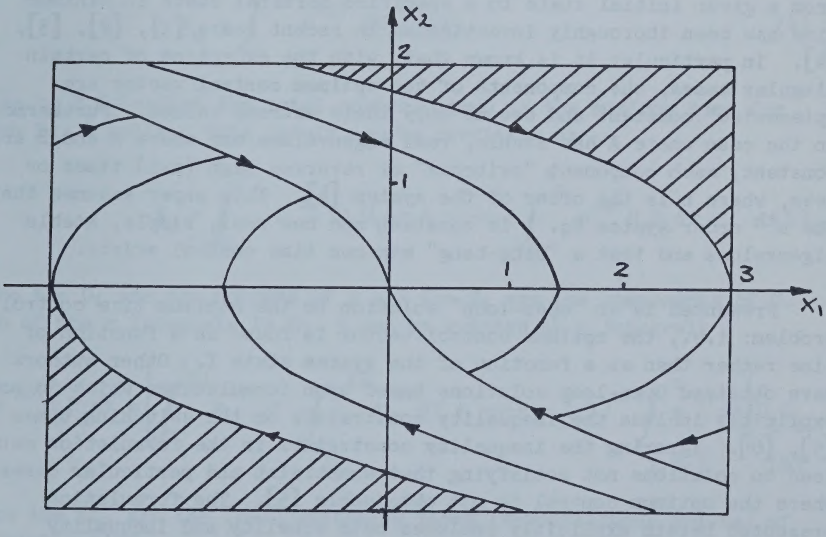


Figure 4
Controllable and Uncontrollable Regions for the $\frac{1}{s^2}$ plant.

INEQUALITY CONSTRAINTS IN THE MINIMUM TIME CONTROL PROBLEM

N. B. HEMESATH
North American Aviation/Autonetics, Anaheim, California

ABSTRACT

This paper develops a novel technique for handling the inequality constraints in the "open-loop" minimum time control problem. The problem is formulated as one of minimizing the total control time, t_n, subject to a set of equality constraints and a set of inequality constraints. The classical Lagrange multiplier technique does not admit inequality constraints and its application, using only the equality constraints, leads to difficulties which are discussed. A simple mathematical artifice allows use of the inequality as well as the equality constraints with Lagrange's method to obtain a normal set of transcendental equations whose solution yields the optimum control.

INTRODUCTION

The problem of determining a control vector, U(t), with magnitude constraints on its r components, which transfers the state of the n^{th} order linear system

$$\dot{X}(t) = A(t)X(t) + B(t)U(t) \tag{1}$$

from a given initial state to a specified terminal state in minimum time has been thoroughly investigated in recent years [1], [2], [3], [4]. In particular it is known that, with the exception of certain singular cases, the components of the optimum control vector are "piecewise" constant and assume only their extreme values. Furthermore in the case where A has stable, real eigenvalues and where A and B are constant, each component "switches" or reverses sign (n-1) times or less, where n is the order of the system [4]. This paper assumes that the n^{th} order system Eq. 1 is constant and has real, simple, stable eigenvalues and that a "bang-bang" minimum time control exists.

Presented is an "open-loop" solution to the minimum time control problem; i.e., the optimum control vector is found as a function of time rather than as a function of the system state X. Other authors have obtained open-loop solutions based upon formulations which do not explicitly include the inequality constraints on the switching times [5], [6]. Ignoring the inequality constraints in the formulation can lead to solutions not satisfying that constraint and particular cases where the optimum control is not obtainable [5]. The formulation presented herein explicitly includes both equality and inequality constraints. It is shown that the optimum control is obtained from a set of transcendental equations.

SWITCHING EQUATIONS

Under the assumptions made above, there exists a nonsingular linear transformation $X = P\underline{X}$ which reduces Eq. 1 to the canonical form

$$\dot{\underline{X}} = \Lambda \underline{X} + DU(t) \qquad (2)$$

where

$$\Lambda = \text{diag}\,(\lambda_1 \text{ ---- } \lambda_n) = P^{-1}AP$$

$$D = P^{-1}B$$

and the λ_i are the eigenvalues of A.

The solution to Eq. 2 is

$$\underline{X}(t) = e^{\Lambda t}\underline{X}_o + \int_o^t e^{\Lambda(t-\tau)}\, DU(\tau)d\tau \qquad (3)$$

If the desired terminal state is the origin, the optimum control vector $U_o(t)$ must satisfy

$$-\underline{X}_o = \int_o^{t_n} e^{-\Lambda\tau}\, DU_o(\tau)d\tau \qquad (4)$$

where $t_n > 0$ (called the total control time) is the minimum time for which Eq. 4 holds. This result can be rewritten as

$$-\underline{X}_o = \int_o^{t_n} e^{-\Lambda\tau}\, D_1U_1 d\tau + \text{----} + \int_o^{t_n} e^{-\Lambda\tau} D_rU_r d\tau \qquad (5)$$

where the D_i are the columns of D and the U_i are the components of U. Each of the r integrals in Eq. 5 can be written as n integrals

$$\int_o^{t_n} e^{-\Lambda\tau} D_iU_i d\tau = \left\{ \int_o^{t_{i1}} e^{-\Lambda\tau} d\tau - \int_{t_{i1}}^{t_{i2}} e^{-\Lambda\tau} d\tau + \text{----} + (-1)^{n-1}\int_{t_{i,n-1}}^{t_n} e^{-\Lambda\tau} d\tau \right\} D_iU_i \qquad (6)$$

where the t_{ij} are the (n-1) switching times (discontinuous points) of U_i which satisfy

$$0 \le t_{i1} \le t_{i2} \le \text{----} \le t_{i,n-1} \le t_n \quad i = 1, \text{--}, r \qquad (7)$$

The integration of Eq. 6, when substituted into Eq. 5, gives a set of n transcendental equations in the variables t_n and t_{ij} $(i = 1, ---, r;\ j = 1, ---, n - 1)$ [7]

$$f_k(t_{ij}, t_n) = 0 \quad k = 1, ---, n \tag{8}$$

The optimum control vector $U_o(t)$ is completely determined by that solution to Eq. 8 which satisfies the constraints in Eq. 7 and for which t_n is minimum. Note that the U_i must also be determined inasmuch as they may take either extreme value. However, assuming that the U_i are known (by trial and error, if no other way) Eq. 8 cannot be solved directly when $r \geq 2$ since there are $r(n - 1) +1$ variables and only n equations.

MINIMIZATION EMPLOYING EQUALITY CONSTRAINTS

Finding the minimum time control from Eq. 8 can be viewed as a constrained minimization problem [5]. Inasmuch as t_n appears in each part of Eq. 8 only once (in a term of the form $k_i e^{-\lambda_i t_n}$), it is possible to solve explicitly for t_n in terms of the remaining $r(n - 1)$ switching times, thus

$$t_n = f(t_{11}, ---, t_{r,n-1}) \tag{9}$$

Substitution of Eq. 9 into the remaining (n - 1) equations of Eq. 8 gives a set of (n - 1) equations independent of t_n

$$g_i(t_{11}, t_{12}, ---, t_{r,n-1}) = 0 \quad i = 1, ---, n - 1 \tag{10}$$

The minimum time control can now be obtained by minimizing $f(t_{11}, t_{12}, ---, t_{r,n-1})$ subject to the (n-1) constraints in Eq. 10 provided the solution satisfies the inequality constraints of Eq. 7.

Using Lagrange multipliers to find the minimum, we form the scalar function

$$H(t_{11}, ---, t_{r,n-1}, a_1, ---, a_{n-1}) = f + a_1 g_1 + --- + a_{n-1} g_{n-1} \tag{11}$$

where the a_i are constant multipliers. Among the stationary points of H, given by

$$\left.\begin{aligned} \frac{\partial H}{\partial t_{ij}} &= 0 \quad i = 1, ---, r \quad j = 1, ---, n - 1 \\ \frac{\partial H}{\partial a_k} &= 0 \quad k = 1, ---, n - 1 \end{aligned}\right\} \tag{12}$$

are the minima of f subject to the constraints $g_i = 0$. Two difficulties are apparent:

1. Solutions obtained may not satisfy the inequality constraints.

2. The desired minimum time solution may fall on the boundary of the closed set defined by the inequality constraints (7) in which case it is not necessarily a stationary point of H and cannot be obtained by this method.

A method which incorporates both equality and inequality constraints is given next.

MINIMIZATION USING BOTH EQUALITY AND INEQUALITY CONSTRAINTS

Through a simple mathematical artifice, the inequality constraints can be incorporated into the formulation when Lagrange multipliers are used. Observe that an inequality can be transformed into an equality by introducing a variable parameter S. For example $f(x) \geq 0$ implies $f(x) = S^2$.

The inequality constraints in Eq. 7 can be rewritten as

$$\begin{aligned} t_{i1} &\geq 0 \\ t_{i2} &\geq t_{i1} \qquad\qquad i = 1, ---, r \\ &\vdots \\ t_n &\geq t_{i,n-1} \end{aligned} \tag{13}$$

and replaced by the following equalities

$$\begin{aligned} t_{i1} &= S_{i1}^2 \\ t_{i2} &= t_{i1} + S_{i2}^2 \qquad\qquad i = 1, ---, r \\ &\vdots \\ t_n &= t_{i,n-1} + S_{in}^2 \end{aligned} \tag{14}$$

Successive substitutions give

$$\begin{aligned} t_{i1} &= S_{i1}^2 \\ t_{i2} &= S_{i1}^2 + S_{i2}^2 \qquad\qquad i = 1, ---, r \\ &\vdots \\ t_n &= S_{i1}^2 + S_{i2}^2 + --- + S_{in}^2 \end{aligned} \tag{15}$$

Now, finding the minimum time control requires minimizing

$$t_n = f(t_{11}, ---, t_{r,n-1})$$

subject to the equality constraints

$$g_k(t_{11}, ---, t_{r,n-1}) = 0 \quad k = 1, ---, n - 1$$

$$\left.\begin{array}{l} t_{i1} = S_{i1}^2 \\ t_{i2} = S_{i1}^2 + S_{i2}^2 \\ \quad \vdots \\ t_n = S_{i1}^2 + S_{i2}^2 + --- + S_{in}^2 \end{array}\right\} i = 1, ---, r \qquad (16)$$

Observe that substituting the relations from Eq. 16, defining the t_{ij} into f and the g_i, leaves r remaining constraint equations in Eq. 16. Thus the problem reduces itself to minimizing

$$t_n = f(S_{ij}) \qquad (17)$$

subject to the constraints

$$\left.\begin{array}{ll} g_k(S_{ij}) = 0 & i = 1, ---, r; \ j = 1, ---, n - 1 \\ h_m(S_{ij}) = -f(S_{ij}) + S_{m1}^2 + --- + S_{mn}^2 = 0 & \\ & k = 1, ---, n - 1 \quad m = 1, ---, r \end{array}\right\} \qquad (18)$$

The minimization can be accomplished by the usual Lagrange technique of forming the scalar function

$$H(S_{ij}; a_1, ---, a_{n-1}; b_1, ---, b_r) = f + \sum_{k=1}^{n-1} a_k g_k + \sum_{m=1}^{r} b_m h_m \qquad (19)$$

where the a_k and b_m are constant multipliers. The stationary points of H are given by

$$\left.\begin{array}{ll} \frac{\partial H}{\partial S_{ij}} = \frac{\partial f}{\partial S_{ij}} + \sum_{k=1}^{n-1} a_k \frac{\partial g_k}{\partial S_{ij}} + \sum_{m=1}^{r} b_m \frac{\partial hm}{\partial S_{ij}} = 0 & \\ \frac{\partial H}{\partial a_k} = g_k = 0 & \begin{array}{l} i = 1, --, r \\ j = 1, --, n \\ k = 1, --, n - 1 \\ m = 1, --, r \end{array} \\ \frac{\partial H}{\partial b_m} = h_m = 0 & \end{array}\right\} \qquad (20)$$

This is a set of (n + 1) (r + 1) - 2 equations in the same number of variables. However, in practice r of these equations may be eliminated immediately by noting that the functions f and g_i are independent of S_{1n}, ---, S_{rn}. Thus

$$\frac{\partial H}{\partial S_{in}} = 2b_i S_{in} = 0 \qquad i = 1, ---, r \tag{21}$$

and it is necessary that either $b_i = 0$ or $S_{in} = 0$. Inasmuch as it is always possible to select the t_{ij} so that

$$t_{i,n-1} < t_n \qquad i = 1, ---, r$$

then $S_{in} \neq 0$, thus implying $b_i = 0$ in Eq. 21. Therefore, the system to be solved always reduces itself to the following normal one of dimension n(r + 1) - 1

$$\left.\begin{aligned} &\frac{\partial H}{\partial S_{ij}} = \frac{\partial f}{\partial S_{ij}} + \sum_{k=1}^{n-1} a_k \frac{\partial g_k}{\partial S_{ij}} \\ &\frac{\partial H}{\partial a_m} = g_m = 0 \qquad\qquad \begin{aligned} i &= 1, ---, r \\ j &= 1, ---, n - 1 \\ m &= 1, ---, n - 1 \\ p &= 1, ---, r \end{aligned} \\ &\frac{\partial H}{\partial b_p} = h_p = 0 \end{aligned}\right\} \tag{22}$$

Among the solutions to this set will be the unique minimum time solution. There are two further conclusions; namely:

1. Every solution (only real solutions considered) is a realizable control because Eq. 16 forces all switching times to be positive and ordered.

2. The minimum time solution will always occur at a stationary point of H inasmuch as the variables S_{ij}, a_k, and b_m are all unrestricted in range.

SUMMARY

A solution to the "open-loop" minimum time control problem has been presented. The optimum control vector satisfies a set of transcendental equations of order n(r + 1) - 1. Inasmuch as the inequality constraints are included in the formulation, all solutions satisfy these constraints, and the optimum control is always a solution to the transcendental set. Although the formulation presented is for time invariant systems with real characteristic roots, it is conceptually correct for systems with complex roots but more difficult to implement because the number of control function sign reversals is not known in general for this case.

REFERENCES

[1] Bellman, R., Glicksberg, I., and Gross, O. "On the Bang-Bang Control Problem," Quarterly Journal of Applied Mathematics, Vol. 14 (1956) pp. 11-18.

[2] LaSalle, J. P. "Time Optimal Control Systems," Proceedings of the National Academy of Sciences, Vol. 45 (1959), pp. 573-577.

[3] DeSoer, C. A. "The Bang-Bang Servo Problem Treated by Variational Techniques," Information and Control, Vol. 2 (1959), pp. 333-348.

[4] Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mischenko, E. F., The Mathematical Theory of Optimal Processes. New York: Interscience Publishers, 1962.

[5] Lee, E. B., "Mathematical Aspects of the Synthesis of Linear Minimum Response-Time Controllers," IRE Professional Group on Automatic Control Transactions, Vol. AC-5 (September 1960), pp. 283-289.

[6] Smith, F. B., Jr. "Time Optimal Control of Higher-Order Systems," IRE Professional Group on Automatic Control Transactions, Vol. AC-6 (February 1961), pp. 16-21.

[7] Hemesath, N. B. "On the Determination of Time Optimal Controls for Linear Stationary Systems," Ph. D thesis, Michigan State University, East Lansing, Michigan, 1964.

OPTIMAL DESIGN OF REGULATORS

N. N. PURI

Re-entry Systems Department
General Electric Company
Philadelphia, Pennsylvania

ABSTRACT

In this paper a simplified algorithm for designing an optimal controller is presented. Plant transfer function $G_p(s)$ is given by:

$$G_p(s) = \frac{N_p(s)}{D_p(s)} = \frac{\sum_{i=0}^{m} b_{i+1} s^i}{s^n + \sum_{j=0}^{n-1} a_{j+1} s^j},$$

where a_i and b_i are constants.

Performance functional I is selected as

$$I = \frac{1}{2} \int_0^\infty \left[\sum_{i=1}^{n} \alpha_i x_i^2 + \sum_{j=1}^{m+1} \beta_i u_j^2 \right] dt$$

where

$$x_i = \frac{d^{i-1}}{dt^{i-1}} x, \quad u_j = \frac{d^{j-1}}{dt^{j-1}}$$

and α_i and β_i are positive constants. The closed loop characteristic equation is obtained as $\Delta(s) = 0$, where

$$\Delta_1(s)\Delta_1(-s) = \left[D_p(s)D_p(-s)\right] \left[\sum_{J=1}^{m+1} \beta_i (-s)^{j-1} (s)^{j-1} \right] +$$

$$\left[N_p(s)\, N_p(-s) \right] \left[\sum_{i=1}^{n} \alpha_i (-s)^{i-1} (s)^{i-1} \right]$$

As a practical example, a three mode controller is discussed in detail where the coefficients k_0, k_1 and k_2 of the three mode controller transfer function $G_c(s) = k_0/s + k_1 + k_2 s$ are selected in optimal fashion.

INTRODUCTION

The usual regulator problem discussed in the literature [1] is:

Given a linear stationary system (Figure 1) in the vectorial form

$$\dot{\underset{\sim}{y}}(t) = \underset{\sim}{A}\underset{\sim}{y}(t) + \underset{\sim}{b}\; u(t) \qquad \underset{\sim}{y}(o) = \underset{\sim}{y}_o \tag{1}$$

$$x(t) = \underset{\sim}{c}'\,\underset{\sim}{y}(t) \tag{2}$$

and

$$W = \underset{\sim}{y}'(t)\,\underset{\sim}{Q}\,\underset{\sim}{y}(t) + u^2(t) \tag{3}$$

find a control law

$$u(t) = \underset{\sim}{d}'\,\underset{\sim}{y}\,(t) \tag{4}$$

such that it minimizes the performance functional

$$I = \int_0^\infty W\,dt \tag{5}$$

The various symbols used above are defined as follows:

$\underset{\sim}{y}(t)$	nx1 state vector
$x(t)$	scalar output
$u(t)$	scalar control input to the plant
$\underset{\sim}{Q}$	n x n positive definite (or semidefinite) matrix
$\underset{\sim}{b}, \underset{\sim}{c}, \underset{\sim}{d}$	n x 1 vectors

Prime denotes the transpose of the matrix or the vector.

The optimal solution takes the form [1, 2]

$$u(t) = -\underset{\sim}{b}'\,\underset{\sim}{P}\,\underset{\sim}{y}(t) = -\underset{\sim}{d}'\,\underset{\sim}{y}(t) \tag{6}$$

where P is a n x n positive definite symmetric matrix which satisfies the Riccati equation

$$\underset{\sim}{P}\underset{\sim}{A} + \underset{\sim}{A}'\,\underset{\sim}{P} - \underset{\sim}{P}\underset{\sim}{b}\underset{\sim}{b}'\,\underset{\sim}{P} + \underset{\sim}{Q} = 0 \tag{7}$$

There are two objections to the above formulation and solution. (1) The Performance Index I is given in terms of the state variable and the control vector u. In reality, Performance Index should be a functional involving controlled output x(t) and its various derivatives and the control function u and its various derivatives. For such a situation the above formulation is inadequate. (2) The control law requires determination of n components of n x 1 vector d. But to obtain these n components, n(n+1)/2 elements of P matrix are determined from equation (7). A better formulation should require solving only n equations.

In what follows, we shall formulate and solve the optimal controller problem in which the Performance Index I is given by

$$I = \int_0^\infty W\,dt$$

where

$$W = \sum_{i=1}^{n} \alpha_i x_i^2 + \sum_{j=1}^{m+1} \beta_j u_j^2 \qquad (9)$$

where

$$x_i = \frac{d^{i-1} x}{dt^{i-1}} \qquad i = 1, \ldots, n$$

$$u_j = \frac{d^{i-1} u}{dt^{i-1}} \qquad j = 1, \ldots, n \qquad (10)$$

$$n \geq m$$

α_i and β_i are non negative constants, x is the output and u is the control input. Notice that such cross product terms as $x_i x_j$ and $u_i u_j$ $(i \neq j)$ are not included in W, because these terms can be integrated, resulting in augmentation of α_i and β_i and changing I by some constant terms depending upon initial conditions only.

In spite of the great advances in the control systems technology, the three mode controller, called the "Proportional plus rate plus Integral" controller is most widely used in practice. The parameters of the controller are selected by such well known methods as the Nichols chart, Bode plot, root locus and Guilliman-Truxal pole zero configuration. Very often the controllers have enough margin of safety so that the field engineer, after going through certain routine procedures is able to tune the controller to the process. All these methods have their advantages and drawbacks depending upon the form in which the specification of the control system are available. In this paper, as an illustration of the above technique, we shall discuss how to select the parameters of a typical three term optimal controller.

STATEMENT OF THE PROBLEM

Given a plant (Figure 2) with transfer function

$$G_p(s) = \frac{N_p(s)}{D_p(s)} = \frac{b_{m+1} s^m + \ldots + b_1}{s^n + a_n s^{n-1} + \ldots + a_1}, \qquad (11)$$

it is required to find a controller

$$G_c(s) = \frac{N_c(s)}{D_c(s)} = \frac{\sum_{i=0}^{n+q-1} k_i s^i}{s^q}, \quad q \leq m \qquad (12)$$

such that the Performance Index

$$I = \int_0^\infty W\,dt = \frac{1}{2}\int_0^\infty \left[\sum_{i=1}^{n} \alpha_i x_i^2 + \sum_{j=1}^{m+1} \beta_j u_j^2\right] dt \tag{13}$$

is minimized. The various symbols used above are defined as:

x system variable to be controlled

$$x_i = \frac{d^{i-1}}{dt^{i-1}} x \qquad (i = 1, \ldots, n)$$

u control variable

$$u_j = \frac{d^{j-1}}{dt^{j-1}} u \qquad (j = 1, \ldots, m) \tag{14}$$

The coefficients a_i, b_i, α_i and β_i are known positive constants

$$x_i(\infty) = 0 \quad \text{(The system is stable.)} \quad (i = 1, \ldots, n)$$

x_i and u_i may take any initial value. Determination of k_i in terms of the known parameters a_i, b_i, α_i and β_i constitutes an optimal controller.

OPTIMAL CONTROL LAW

From Parseval's Theorem, the Performance Index I can be written as

$$I = \left[\frac{1}{2 \times 2\pi j}\int_{c-j\infty}^{c+j\infty} F(s)\,x(s)\,x(-s)\,ds\right] + \text{terms due to initial conditions on } x_i,\ u_i, \text{ etc.} \tag{15}$$

where

$$F(s) = \left[\alpha_1 - \alpha_2 s^2 \ldots (-1)^{n-1}\alpha_n s^{2(n-1)}\right] + \left[\beta_1 - \beta_2 s^2 \ldots (-1)^m \beta_{m+1} s^{2m}\right] \times \frac{D_p(s)\,D_p(-s)}{N_p(s)\,N_p(-s)} \tag{16}$$

or

$$F(s)\,N_p(s)\,N_p(-s) = \left[N_p(s)\,N_p(-s)\right]\left[\sum_{i=1}^{n} (-1)^{i-1}\alpha_i s^{2(i-1)}\right] + \left[D_p(s)\,D_p(-s)\right]\left[\sum_{j=1}^{m+1} (-1)^{j-1}\beta_j s^{2(j-1)}\right] \tag{17}$$

Let

$$\Delta(s) = F(s)\, N_p(s)\, N_p(-s) = \Delta_1(s)\, \Delta_1(-s) \tag{18}$$

where $\Delta_1(s)$ is a polynomial which has all its roots in the left half (stable half) of the s plane. We will now state the optimal control law and prove it in the next Section.

"Optimal control function u(t) is such that $\Delta_1(s)$ in (18) represents the closed loop characteristic polynomial."

Thus, if the transfer function of the controller is $N_c(s)/D_c(s)$, then (Fig. 3A)

$$\Delta_1(s) = D_c(s)\, D_p(s) + N_c(s)\, N_p(s) \tag{19}$$

If the degree of $\Delta_1(s)$ is n+q ($q \leq m$), then, as in (12)

$$D_c(s) = s^q$$

$$N_c(s) = \sum_{i=0}^{n+q-1} k_i\, s^i \tag{20}$$

Equating coefficients on each side of (19), n+q coefficients $k_o, \ldots, k_{n+q-1}$ are determined. This solves the optimal control problem. In the next section, we will prove the optimal control law.

PROOF OF OPTIMAL CONTROL LAW

From Figure 2

$$x(s) = \frac{N_p(s)}{D_p(s)}\, U(s) = \frac{b_{m+1}\, s^m + \ldots + b_1}{s^n + q_u\, s^{n-1} + \ldots + a_1}\, U(s) \tag{21}$$

or

$$x^{(n)} + a_n\, x^{(n-1)} + \ldots + a_1\, x = b_{m+1}\, u^{(m)} + \ldots + b_1\, u \tag{22}$$

where

$$x^{(n)} = \frac{d^n x}{dt^n}$$

$$u^{(m)} = \frac{d^m u}{dt^m} \tag{23}$$

Let

$$\dot{x}_i = x_{i+1} \qquad i = 1, \ldots, n-1$$

$$\dot{u}_j = u_{j+1} \qquad j = 1, \ldots, m-1 \tag{24}$$

Equation (23) can be represented by n+m-1 single order differential equations

$$
\begin{aligned}
g_1 &= \dot{x}_1 - x_2 = 0 \\
g_2 &= \dot{x}_2 - x_3 = 0 \\
&\vdots \\
g_n &= \dot{x}_n + q_n x_n + \dots + a_1 x_1 - b_1 u_1 \dots - b_m u_m - b_{m+1} \dot{u}_m \\
g_{n+1} &= \dot{u}_1 - u_2 \\
&\vdots \\
g_{n+m-1} &= \dot{u}_{m-1} \; u_m
\end{aligned}
\tag{25}
$$

Thus, it is required that the functional I be minimized subject to constraint (25). This is a well known variational problem. Essentially we introduce a new function

$$
V = W + \psi_1 g_1 + \dots + \psi_n g_n + \lambda_1 g_{n+1} \dots \lambda_{m-1} g_{n+m-1}
\tag{26}
$$

where $\psi_1, \dots, \psi_n$ and $\lambda_1, \dots, \lambda_{m-1}$ are referred to as Lagrange multipliers or co-state variables and they are functions of time.

From (25) and (13), the equation (26) can be rewritten as

$$
\begin{aligned}
V = \frac{1}{2}\left[\sum_{i=1}^{n} \alpha_i x_i^2 + \sum_{j=1}^{n} \beta_i u_m^2 + \beta_{m+1} \dot{u}_m^2\right] &+ \psi_1 (\dot{x}_1 - x_2) + \dots \\
&+ \psi_n (\dot{x}_n + a_n x_n + \dots + a_1 x_1 - b_1 u_1 - \dots - b_m u_m - b_{m+1} \dot{u}_m) \\
&+ \lambda_1 (\dot{u}_1 - u_2) + \dots + \lambda_{m-1} (\dot{u}_{m-1} - u_m)
\end{aligned}
\tag{27}
$$

The variational equations are

$$
\begin{aligned}
&\frac{d}{dt}\left[\frac{\partial V}{\partial \psi_i}\right] = \frac{\partial V}{\partial \psi_i} &&& \frac{d}{dt}\left[\frac{\partial V}{\partial \lambda_j}\right] &= \frac{\partial V}{\partial \lambda_j} \\
&&& i = 1, \dots, n \\
&\frac{d}{dt}\left[\frac{\partial V}{\partial \dot{x}_i}\right] = \frac{\partial V}{\partial x_i} &&& \frac{d}{dt}\left[\frac{\partial V}{\partial \dot{u}_j}\right] &= \frac{\partial V}{\partial u_j} \qquad j = 1, \dots, m-1 \\
&&&& \frac{d}{dt}\left[\frac{\partial V}{\partial \dot{u}_m}\right] &= \frac{\partial V}{\partial u_m}
\end{aligned}
\tag{28}
$$

This results in the following additional differential equations,

$$\dot\psi_1 = \qquad + a_1\psi_n + \alpha_1 x_1$$

$$\dot\psi_2 = -\psi_1 \qquad + a_2\psi_n + \alpha_2 x_2$$

$$\dot\psi_3 = -\psi_2 \qquad + a_3\psi_n + \alpha_3 x_3$$

$$\vdots$$

$$\dot\psi_n = \qquad -\psi_{n+1} + a_n\psi_n + \alpha x_n \tag{29}$$

$$\dot\lambda_1 = \qquad - b_1\psi_n + \beta_1 u_1$$

$$\dot\lambda_2 = -\lambda_1 \qquad - b_2\psi_n + \beta_2 u_2$$

$$\vdots$$

$$\dot\lambda_{m-1} = \qquad -\lambda_{m-2} - b_{m-1}\psi_n + \beta_{m-1} u_{m-1} \tag{30}$$

$$\beta_{m+1}\ddot u_m = -\lambda_{m-1} + b_{m+1}\psi_n - b_m\psi_n + \beta_m u_m \tag{31}$$

Equations (29) can be combined together to yield an nth order differential equation in ψ_n.

Thus

$$\left[\psi_n^{(n)} - a_n\psi_n^{(n-1)} + \ldots + (-n)^n a_1\psi_n\right] = \left[\alpha_n x_n^{(n-1)} - \alpha_{n-1}{}^{(n-2)} \cdots \right.$$

$$\left. (-1)^{(n-1)}\alpha_1 x_1\right] \tag{32}$$

But

$$x_i^{(j)} = x^{(i+j-1)} \tag{33}$$

Thus

$$\psi_n(s) = \frac{-\left[\alpha_1 - \alpha_2 s^2 \ldots (-1)^{n-1}\alpha_n s^{2(n-1)}\right]}{\left[a_1 + a_2(s) + \ldots + a_n(-s)^{n-1} + (-s)^n\right]} X(s) + \text{terms due to initial conditions.} \tag{34}$$

Similarly equations (30) can be combined together to yield a $(m-1)^{th}$ differential equation

$$\lambda_{m-1}^{(m-1)} = \left[-b_{m-1}\,\psi_n^{(m-2)} + b_{m-2}\,\psi_n^{(m-3)} \ldots (-1)^{m-1}\,\psi_n\right] - \left[-\beta_{m-1}\,u_{m-1}^{(m-2)} + \beta_{m-2}\,u_{m-2}^{(m-3)} \ldots (-1)^{m-1}\,u_1\right] \tag{35}$$

Combining (31) and (35) yields

$$\left[\beta_{m+1}\,u_m^{(m+1)} - \beta_m\,u_m^{m-1} + \beta_{m-1}\,u_{m-1}^{(m-2)} \ldots (-1)^m\,\beta_1\,u_1\right] = \left[b_{m+1}\,\psi_n^{m} - b_m\,\psi_n^{m-1} \ldots (-1)^m\,b_1\,\psi_n\right] \tag{36}$$

or

$$U(s) = \frac{N_p(-s)}{\sum\limits_{j=1}^{m+1} (-1)^{j-1}\beta_j\, s^{2(j-1)}}\,\psi_n(s) + \text{terms due to initial conditions} \tag{37}$$

From (34) and (37)

$$U(s) = -\frac{\left[\sum\limits_{i=1}^{n} (-1)^{i-1}\alpha_i\, s^{2(i-1)}\right]}{\left[\sum\limits_{j=1}^{m+1} (-1)^{j-1}\beta_j\, s^{2(j-1)}\right]}\,\frac{N_p(-s)}{D_p(-s)}\,X(s) + \text{terms due to initial conditions} \tag{38}$$

But

$$\frac{U(s)}{X(s)} = \frac{D_p(s)}{N_p(s)}$$

Thus

$$\left[1 + \frac{\left(\sum\limits_{i=1}^{n} (-1)^{i-1}\alpha_i\, s^{2(i-1)}\right)}{\left(\sum\limits_{j=1}^{m+1} (-1)^{j-1}\beta_j\, s^{2(j-1)}\right)}\,\frac{N_p(-s)\,N_p(s)}{D_p(-s)\,D_p(s)}\right] U(s) = \text{terms due to initial conditions} \tag{39}$$

Hence the characteristic determinant for system of equations (25), (29), (30) and (31) is given by

$$\Delta(s) = \Delta_1(s)\,\Delta_1(-s) = \left(D_p(s)\,D_p(-s)\right)\left(\sum_{j=1}^{m+1} (-1)^{j+1}\beta_j\, s^{2(j-1)}\right) + \left(N_p(s)\,N_p(-s)\right)\left(\sum_{i=1}^{n} (-1)^{i-1}\alpha_i\, s^{2(i-1)}\right) \tag{40}$$

Equation (40) shows that corresponding to every root in the left half plane there is an equal and opposite root in the right half plane. A root at $s = s_i$ contributes a root of the form $e^{s_i t}$ in the response of the system variable x (or u). Since at the outset we assume the system to be stable, the contribution to the system response due to right half plane roots should be zero. Thus the final Control System characteristic polynomial should include only roots in the left half plane. Hence the roots of $\Delta_1(s)$ are the same as the roots of the closed loop system characteristic equation. This completes the proof.

Three Mode Controller

Three mode controller is an excellent example of application of the above technique. Actual derivation of plant transfer function $G_p(s)$ involves various simplifying approximations such as neglecting very small time constants, or the pole zero pairs located near the $j\omega$ axis. A reasonably good model for some processes, which retains the critical parameter, may be represented by Figure 3.

$$G_p(s) = \frac{b_2 s + b_1}{s^2 + a_2 s + a_1} \tag{41}$$

or

$$\ddot{x} + a_2 \dot{x} + a_1 x = b_2 \dot{u} + b_1 u \tag{42}$$

It is required to find a three mode controller

$$G_c(s) = \frac{k_0}{s} + k_1 + k_2 s \tag{43}$$

Such that the performance index

$$I = \int_0^\infty \left(\frac{\alpha_1 x^2}{2} + \frac{\alpha_2 \dot{x}^2}{2} + \frac{\dot{u}^2}{2} \right) dt \tag{44}$$

is minimized.

From Figure 3

$$\ddot{x} + a_2 \dot{x} + a_1 x = b_2 \dot{u} + b_1 u \tag{45}$$

Set

$$\dot{x} = x_1$$

$$\dot{x}_1 = x_2 \tag{46}$$

Then

$$g_1 = \dot{x}_1 - x_2 = 0$$

$$g_2 = \dot{x}_2 + a_2 x_2 + a_1 x_1 - b_2 \dot{u} - b_1 u = 0 \tag{47}$$

$$V = \frac{\alpha_1 x_1^2}{2} + \frac{\alpha_2 x_2^2}{2} + \frac{\dot{u}^2}{2} + \psi_1 (\dot{x}_1 - x_2) +$$

$$\psi_2 (\dot{x}_2 + a_2 x_2 + a_1 x_1 - b_2 \dot{u} - b_1 u) \tag{48}$$

The variational equations are

$$\left| \frac{d}{dt} \left(\frac{\partial V}{\partial \dot{\psi}_i} \right) = \frac{\partial}{\partial \psi_i} V \right. \tag{49}$$

$$\frac{d}{dt} \left(\frac{\partial V}{\partial \dot{x}_i} \right) = \frac{\partial}{\partial x_i} V \qquad i = 1,\ 2 \tag{50}$$

$$\frac{d}{dt} \left(\frac{\partial V}{\partial \dot{u}} \right) = \frac{\partial}{\partial u} V \tag{51}$$

but

$$\frac{\partial V}{\partial \psi_1} = \dot{x}_1 - x_2 \qquad \frac{\partial V}{\partial \dot{\psi}_1} = 0$$

$$\frac{\partial V}{\partial \psi_2} = \dot{x}_2 + a_2 x_2 + a_1 x_1 - b_2 \dot{u} - b_1 u \qquad \frac{\partial V}{\partial \dot{\psi}_2} = 0$$

$$\frac{\partial V^2}{\partial x_1} = a_1 \psi_2 + \alpha_1 x_1 \qquad \frac{\partial V}{\partial \dot{x}_1} = \psi_1 \tag{52}$$

$$\frac{\partial V}{\partial x_2} = -\psi_1 + a_2 \psi_2 + \alpha_2 x_2 \qquad \frac{\partial V}{\partial \dot{x}_2} = \psi_2$$

$$\frac{\partial V}{\partial u} = -b_1 \psi_2 \qquad \frac{\partial V}{\partial \dot{u}} = -b_2 \psi_2 + \dot{u}$$

Thus

$$\dot{x}_1 = +x_2$$

$$\dot{x}_2 = -a_1 x_1 - a_2 x_2 - b,u - b_2 \dot{u}$$

$$\dot{\psi}_1 = + \alpha_1 x_1 + a_1 \psi_2 \tag{53}$$

$$\dot{\psi}_2 = + \alpha_2 x_2 - \psi_1 + a_2 \psi_2$$

$$\ddot{u} = -b_1 \psi_2 + b_2 \dot{\psi}_2$$

Eliminating $\dot{\psi}_2$ and letting

$$u = u_1$$

$$\dot{u} = u_2 \tag{54}$$

we obtain a matrix relation

$$\underset{\sim}{\dot{z}} = \underset{\sim}{A}\,\underset{\sim}{z} \tag{55}$$

where $\underset{\sim}{z}$ is an n dimensional vector and $\underset{\sim}{A}$ is an nxn matrix, given as

$$\underset{\sim}{z} = \begin{bmatrix} x_1 \\ x_2 \\ \psi_1 \\ \psi_2 \\ u_1 \\ u_2 \end{bmatrix}, \quad \underset{\sim}{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -a_1 & -a_2 & 0 & 0 & b_1 & b_2 \\ \alpha_1 & 0 & 0 & a_1 & 0 & 0 \\ 0 & \alpha_2 & -1 & a_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & b_2\,\alpha_2 & -b_2 & a_2 b_2 - b_1 & 0 & 0 \end{bmatrix} \tag{56}$$

Solution of the Controller Parameters

The characteristic determinant of the optimal system variational equation is given by

$$\Delta\,(s) = \left| s\,\underset{\sim}{I} - \underset{\sim}{A} \right| = 0 \tag{57}$$

Where $\underset{\sim}{I}$ is identity matrix. This determinant contains only even powers of s and can be written as

$$\Delta\,(s) = s^6 - r_1 s^4 + r_2 s^2 - r_3 \tag{58}$$

where

$$\begin{aligned} r_1 &= -2a_1 + a_2^2 + \alpha_2 b_2^2 \\ r_2 &= a_1^2 + \alpha_2 b_1^2 + \alpha_1 b_2^2 \\ r_3 &= +\alpha_1 b_1^2 \end{aligned} \tag{59}$$

Substituting $s^2 = \lambda$, in (58), yields

$$\Delta\,(\lambda) = \lambda^3 - r_1\,\lambda^2 + r_2\,\lambda - r_3 \tag{60}$$

As long as r_3 is positive, which is the situation in present case, the equation (26) will have one real positive root and a pair of complex conjugate roots, as shown in figure 4, a, b and c. Any of these three distributions of the roots can occur depending upon the values of r_1, r_2 and r_3. The resultant root distribution in the s-plane is such that we have two equal and opposite real roots and a pair each of complex roots in each of the half planes, as shown figure 5. Now each root at $s = s_j$, of figure 5, contributes a term of the form $e^{s_i t}$ to the response of the system variable x. Since the system is stable $(x(\infty) = \dot{x}(\infty) = 0)$, the contribution of the poles in the right half plane should be zero. Therefore the final control system determinant should include only roots in the left half plane.

Thus if the determinant Δ (s) in equation (26) is written as

$$\Delta(s) = -\Delta_1(s)\,\Delta_1(-s) \tag{61}$$

Then $\Delta_1(s)$, is the optimal control system determinant and contains only the roots in the left half plane. The coefficients of various powers of s in it are real. $\Delta_1(s)$ can be found either by direct location of the roots of Δ (s) or in the following manner:

Let

$$\Delta(s) = s^3 + p_1 s^2 + p_2 s + p_3 \tag{62}$$

Thus multiplying $\Delta_1(s)$ with $(-\Delta_1(-s))$ and equating coefficient with Δ (s), we have

$$r_1 = p_1^2 - 2p_2 \tag{63}$$

$$r_2 = p_2^2 - 2p_1 p_3 \tag{64}$$

$$r_3 = p_3^2 \tag{65}$$

Thus

$$p_3 = \sqrt{r_3} \tag{66}$$

and

$$p_1^4 - 2 r_1 p_1^2 - 8\sqrt{r_3}\, p_1 + (r_1^2 - 4r_2) = 0 \tag{67}$$

The largest real positive root of (68) gives the value of p_1 to be used here. The remaining coefficient p_2 is obtained from (63) and the determination of Δ (s) is complete. Bayard's method [5] is useful for high order polynomials.

Now the close loop transfer function of the system in figure 3 is

$\dfrac{G_c(s)G_p(s)}{1 + G_c(s)G_p(s)}$ The characteristic equation of the close loop system is

$$1 + G_{c(s)}\,G_{p(s)} = 0 \tag{68}$$

or

$$1 + \frac{(k_0 + k_1 s + k_2 s^2)(b_2 s + b_1)}{s(s^2 + a_2 s + a_1)} \tag{69}$$

or

$$s^3 + \frac{(a_2 + b_1 k_2 + b_2 k_1)}{1 + b_2 k_2} s^2 + \frac{(a_i + k_1 b_1 + k_o b_2)}{1 + b_2 k_2} s + \frac{k_o b_1}{1 + b_2 k_2} = 0 \qquad (70)$$

Equating the left hand side of (38) with Δ (s), we have

$$\begin{bmatrix} 0 & b_2 & -p_1 b_2 + b_1 \\ b_2 & b_1 & -p_2 b_2 \\ b_1 & 0 & -p_3 b_2 \end{bmatrix} \begin{bmatrix} k_o \\ k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} -\alpha_2 + p_1 \\ -a_1 + p_2 \\ p_3 \end{bmatrix} \qquad (71)$$

or

$$k_o = \frac{1}{\Delta} \frac{p_3 (a_2 b_1 b_2 - a_1 b_2^2 - b_1^2)}{}$$

$$k_1 = \frac{1}{\Delta} \frac{a_1 b_1^2 - a_1 b_1 b_2 p_1 + (a_2 b_1 b_2 - b_1^2) p_2 + (b_1 b_2 - a_2 b_2^2) p_2}{} \qquad (72)$$

$$k_2 = \frac{1}{\Delta} \frac{(-a_2 b_1^2 + a_1 b_1 b_2 + b_1^2 p_1 - b_1 b_2 p_2 - b_2^2 p_3)}{}$$

where

$$\Delta = (b_2^2 p_3 + b_1^2 p_1 - p_2 b_1 b_2)\, b_2 - b_1^3 \qquad (73)$$

CONCLUSION

In this paper a technique is presented for the design of an optimal controller for a plant whose transfer function is ratio of two polynomials. This method avoids the Solution of Riccati equation and essentially involves factorization of a polynomial in s^2 into two factors involving roots in the left half and right half plane. Design of a three mode controller is discussed in detail. Performance index chosen for design is very general and involves the output variable, control variable and their derivatives.

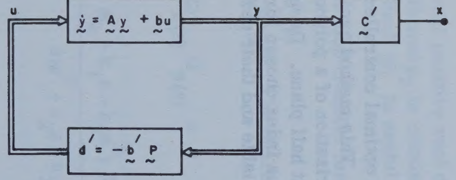

u
ẏ = Ay + bu
y
C′
x
d′ = −b′ P

Figure 1

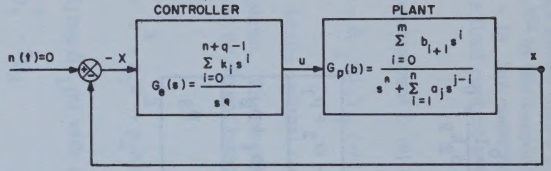

CONTROLLER
PLANT
n (t)=0
− X
G_e(s) = Σ_{i=0}^{n+q−1} k_i s^i / s^q
u
G_p(b) = Σ_{i=0}^{m} b_{i+1} s^i / (s^n + Σ_{j=1}^{n} a_j s^{j−1})
x

Figure 2

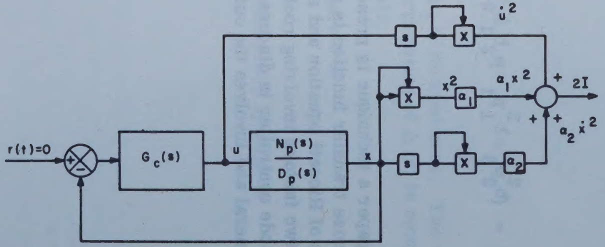

r(t)=0
G_c(s)
u
N_p(s)/D_p(s)
x
s
X
u̇²
x²
a₁
a₁x²
+ 2I
+
+
a₂ẋ²
a₂

Figure 3a

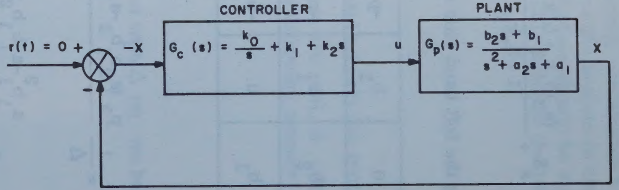

CONTROLLER
PLANT
r(t) = 0 +
−X
−
G_c (s) = k₀/s + k₁ + k₂s
u
G_p(s) = (b₂s + b₁)/(s² + a₂s + a₁)
X

Figure 3b

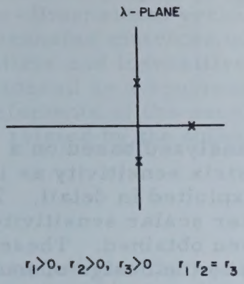


Figure 4a

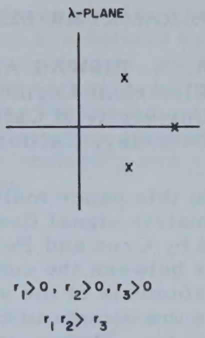


Figure 4b

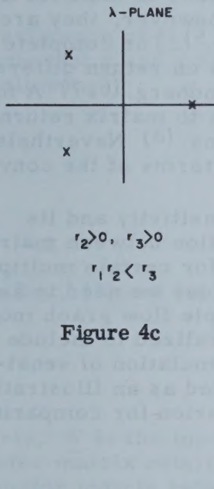
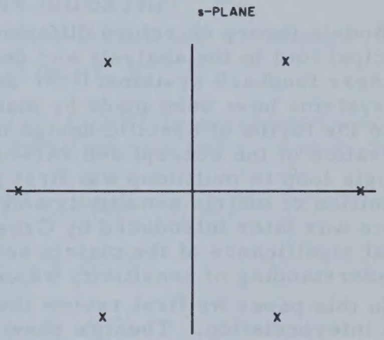


Figure 4c



Root location of $s^6 - r_1 s^4 + r_2 s^2 - r_3 = 0$

Figure 5

REFERENCES

1. Kalman R. E. "When is a Linear Control System Optimal" RIAS TR63-5, March 1963.
2. A. M. Letov; Analytical Design of Regulator I, II and III, Automation and Telemechanic, 21, Nos. 4, 5, 6 (1960).
3. Kalman R. E., Theory of Optimal Control and the Calculus of Variation, Chapter 16, Mathematical Optimization Techniques (1963), Edited by R. Bellman.
4. A. G. J. Macfarlane; - An Eigen-vector solution of the optimal Linear Regulator Problem, Journal of Electronics and Control, pp 643-654, 1962.
5. D. F. Tuttle, Jr., Network Synthesis Vol 1. (Book) John Wiley and Sons Inc., New York pp 1125.
6. Kvasovski, N. N. and Letov, A. M. "Towards the Theory of Analytical Design of Regulators" Automation and Remote Control, Vol 23, No. 6., 1962, pp 649-656.

MULTIPARAMETER SENSITIVITY ANALYSIS FOR LINEAR SYSTEMS

R. N. BISWAS AND E. S. KUH
Electrical Engineering Department
University of California
Berkeley, California

ABSTRACT

In this paper multiparameter system is analyzed based on a simple matrix signal flow graph model. The matrix sensitivity as introduced by Cruz and Perkins is reviewed and exploited in detail. The relations between the conventional multiparameter scalar sensitivity and the elements of the sensitivity matrix are then obtained. These formulas are simple to compute and give insight to the design of multiloop systems. Higher order multiparameter sensitivities are also derived. As an illustration, a complete formulation of sensitivity analysis for general RLC network is presented.

INTRODUCTION

Bode's theory on return difference and sensitivity constitutes the principal tool in the analysis and design of single loop, single parameter linear feedback systems.(1-3) Some extensions to multivariable control systems have been made by many people; however, they are mainly in the forms of specific design methods.(4, 5) The complete generalization of the concept and various theorems on return difference from single loop to multiloop was first given by Sandberg.(6, 7) A formal definition of matrix sensitivity and its relation to matrix return difference was later introduced by Cruz and Perkins.(8) Nevertheless, the actual significance of the matrix sensitivity in terms of the conventional understanding of sensitivity was not clear.

In this paper we first review the matrix sensitivity and its physical interpretation. Then we show the connection between matrix sensitivity and the conventional scalar sensitivity for certain multiparameter, multiloop systems. To derive these relations we need to assume that the system can be represented by a simple flow graph model. We then show that the assumed model can be generalized to include other complicated configurations. A complete formulation of sensitivity analysis for general RLC network is presented as an illustration. The paper is concluded with comments on the criterion for comparing systems with regard to their overall sensitivity.

THE SENSITIVITY MATRIX

For all multiparameter, multiloop systems we can always select a set of elements which are sensitive with respect to the overall system. This set can include either elements which are inherently sensitive to variation or elements whose effect on the over-all-performance of the system is of main interest to designers. The remaining elements in the system thus belong to an insensitive set. In conventional electronic circuits, active devices, whose parameters are more susceptible to temperature change and manufacturing tolerance, are considered to be in the sensitive set, while passive components are grouped in the insensitive set.

The research reported here was supported in part by the Joint Services Electronics Program (Air Force Office of Scientific Research, Army Research Office, Office of Naval Research) under Grant AF-AFOSR-139-64 and by the National Science Foundation under Grant GP 2684.

Consider a multiparameter system as represented by the simple matrix flow graph in Fig. 1. The input and output are defined by the finite-dimensional vectors $\underline{u}$ and $\underline{z}$ in the frequency domain. $\underline{X}$ and $\underline{C}$ are transfer matrices which are described by the elements in the sensitive and insensitive sets, respectively. The vector $\underline{v}$ may be considered as a spurious input due to noise or distortion generated in the elements of the sensitive set. The two node vectors $\underline{y}$ and $\underline{z}$ are then related by the following two linear equations:

$$\underline{y} = \underline{C}\,\underline{z} + \underline{u} \tag{1}$$

$$\underline{z} = \underline{X}\,\underline{y} + \underline{v} \tag{2}$$

We assume that the equations have a unique solution for $\underline{y}$ and $\underline{z}$ in terms of inputs $\underline{u}$ and $\underline{v}$. Thus eliminating $\underline{y}$ between Eqs. (1) and (2), we obtain

$$\underline{z} = (\underline{1} - \underline{X}\,\underline{C})^{-1}\,\underline{X}\,\underline{u} + (\underline{1} - \underline{X}\,\underline{C})^{-1}\,\underline{v} \tag{3}$$

Similarly, eliminating $\underline{z}$ between Eqs. (1) and (2), we obtain

$$\underline{y} = (\underline{1} - \underline{C}\,\underline{X})^{-1}\,\underline{u} + (\underline{1} - \underline{C}\,\underline{X})^{-1}\,\underline{C}\,\underline{v} \tag{4}$$

The existence of $(\underline{1} - \underline{X}\,\underline{C})^{-1}$ and $(\underline{1} - \underline{C}\,\underline{X})^{-1}$ is guaranteed since we assume that $\underline{z}$ and $\underline{y}$ are uniquely specified by $\underline{u}$ and $\underline{v}$. Let us define the following

$$\underline{W} = (\underline{1} - \underline{X}\,\underline{C})^{-1}\,\underline{X} = \underline{X}(\underline{1} - \underline{C}\,\underline{X})^{-1} \tag{5}$$

$$\underline{S} = (\underline{1} - \underline{X}\,\underline{C})^{-1} \tag{6}$$

and

$$\underline{\widehat{S}} = (\underline{1} - \underline{C}\,\underline{X})^{-1} \tag{7}$$

Then Eqs. (3) and (4) become

$$\underline{z} = \underline{W}\,\underline{u} + \underline{S}\,\underline{v} \tag{8a}$$

$$\underline{y} = \underline{\widehat{S}}\,\underline{u} + \underline{\widehat{S}}\,\underline{C}\,\underline{v} \tag{8b}$$

Clearly, $\underline{W}$ is the input-output transfer matrix of the system, $\underline{S}$ is a transfer matrix relating the output $\underline{z}$ and the spurious input $\underline{v}$, $\underline{\widehat{S}}$ is a transfer matrix relating the signal $\underline{y}$ and the input $\underline{u}$.

Next we proceed to obtain an expression for the increment $\delta\underline{W}$ in the transfer matrix $\underline{W}$ when $\underline{X}$ changes to $\underline{X}^* = \underline{X} + \delta\underline{X}$. It is easy to show that the increment $\delta\underline{W}$ can be expressed in two equivalent forms:

$$\delta\,\underline{W} = (\underline{1} - \underline{X}^*\,\underline{C})^{-1}\,\delta\underline{X}\,(\underline{1} - \underline{C}\,\underline{X})^{-1} \tag{9}$$

and

$$\delta\,\underline{W} = (\underline{1} - \underline{X}\,\underline{C})^{-1}\,\delta\,\underline{X}(\underline{1} - \underline{C}\,\underline{X}^*)^{-1} \tag{10}$$

In view of the foregoing interpretations of the matrices $\underline{S} = (\underline{1} - \underline{X}\,\underline{C})^{-1}$ and $\underline{\widehat{S}} = (\underline{1} - \underline{C}\,\underline{X})^{-1}$, it is clear that Eqs. (9) and (10) enable us to represent the effect of perturbations in $\delta\underline{X}$ on the input-output relationship of the system by the signal flow graph models shown in Figs. 2a and 2b. Each of these two representations consists of two models of the system, one perturbed and the other unperturbed. The first order perturbations may be represented by $\delta\underline{W}^{o}$ according to

$$\begin{aligned}\delta\underline{W}^{o} &= (\underline{1} - \underline{X}\,\underline{C})^{-1}\,\delta\underline{X}(\underline{1} - \underline{C}\,\underline{X})^{-1} \\ &= \underline{S}\,\delta\underline{X}\,\underline{\widehat{S}} \qquad (11)\end{aligned}$$

and the flow graph representation then consists of two identical models of the unperturbed system. This particular result often proves useful in measuring first order perturbation without making actual changes in the system parameters.

We now consider a more special case, where $\underline{X}$ in Fig. 1 is a nonsingular matrix. From Eq. (5), we have

$$\underline{W}^{-1} = \underline{X}^{-1} - \underline{C}$$

Equation (9) can now be written in the following form:

$$(\delta\underline{W})\underline{W}^{-1} = \underline{S}^{*}(\delta\underline{X})\underline{X}^{-1} \qquad (12)$$

where

$$\underline{S}^{*} = (\underline{1} - \underline{X}^{*}\underline{C})^{-1} \qquad (13)$$

Equation (12) is obviously reminiscent of the definition of conventional scalar sensitivity, and, as such, may be used to define the sensitivity matrix.(8) The sensitivity matrix is therefore given by $\underline{S}^{*}$, the perturbed transfer matrix relating the output vector $\underline{z}$ to a signal $\underline{v}$ applied at the output node. Note that this definition of the sensitivity matrix does not assume small perturbations as in conventional scalar sensitivity. However, if $\delta\underline{X}\to\underline{0}$, the first order sensitivity matrix is given by

$$\underline{S} = (\underline{1} - \underline{X}\,\underline{C})^{-1} = \underline{W}\,\underline{X}^{-1} \qquad (14)$$

$(\underline{1} - \underline{X}\,\underline{C})$ is clearly the return difference matrix at the output node; thus the first order sensitivity matrix is equal to the inverse of the return difference matrix.

Similarly, if Eq. (10) is used, we obtain

$$\underline{W}^{-1}(\delta\underline{W}) = \underline{X}^{-1}(\delta\underline{X})\underline{\widehat{S}}^{*}, \qquad (15)$$

where

$$\underline{\widehat{S}}^{*} = (\underline{1} - \underline{C}\,\underline{X}^{*})^{-1}, \qquad (16)$$

which is the perturbed transfer matrix relating the vector $\underline{y}$ appearing at the input node to the input $\underline{u}$. The first order sensitivity matrix is then given by

$$\underline{\widehat{S}} = (\underline{1} - \underline{C}\,\underline{X})^{-1} = \underline{X}^{-1}\underline{W}, \qquad (17)$$

which is equal to the inverse of the return difference matrix at the input node.

The elements of the three matrices $\underline{W}$, $\underline{S}$, and $\underline{\widehat{S}}$ have the following physical meanings: Let w_{ij}, s_{ij} and $\widehat{s}_{ij}$ be the elements on the i-th row and j-th column of $\underline{W}$, $\underline{S}$ and $\underline{\widehat{S}}$, respectively. Then, with a single unit input $u_j (u_k=0,\ k\neq j)$, the output z_i and y_i represent w_{ij} and $\widehat{s}_{ij}$, respectively. While a single unit input $v_j (v_k=0,\ k\neq j)$, the output z_i represents s_{ij}. For the case where $\underline{X}$ is diagonal, we can obtain further useful results as follows:

Let

$$\underline{X} = \text{diag}\,(x_1,\ x_2 \ldots\ x_n) \tag{18}$$

Then from Eqs. (14) and (17), we have

$$w_{ij} = s_{ij} x_j = x_i \,\hat{s}_{ij} \ . \tag{19}$$

The element of $\delta \underline{W}^o$, which represents the first order perturbation, can then be written from Eq. (11) as

$$\delta w_{ij}^o = \sum_{k=1}^{n} s_{ik}\ \delta x_k\ \hat{s}_{kj} \tag{20}$$

$$= \sum_{k=1}^{n} \frac{w_{ik} w_{kj}}{x_k} \cdot \frac{\delta x_k}{x_k} \tag{21}$$

With the above formulas in mind we can derive the relationship between the elements of the sensitivity matrix and the conventional multiparameter scalar sensitivities. These relationships to be obtained in the following section are in turn useful in the computation and measurement for a complete sensitivity analysis.

THE SCALAR SENSITIVITY COEFFICIENTS

Consider the scalar transfer function w_{ij}, which is a function of n sensitive parameters, x_k, k=1, 2, . . . , n. The multivariable Taylor's series expansion for w_{ij} is

$$\delta w_{ij} = \sum_{k=1}^{n} \frac{\partial w_{ij}}{\partial x_k}\ \delta x_k + \sum_{k,\ell=1}^{n} \frac{\partial^2 w_{ij}}{\partial x_k \partial x_\ell}\ \frac{\delta x_k\, \delta x_\ell}{2!} + \ldots \tag{22}$$

The first term of the expansion gives the first order change of w_{ij} and is equal to $\delta w_{ij}{}^0$ of Eq. (21). Thus

$$\frac{\partial w_{ij}}{\partial x_k} = \frac{w_{ik} w_{kj}}{x_k^2} \tag{23}$$

For a single-input, single-output system, with the input applied at the input node of the j-th sensitive element and the output taken from the output node of the i-th sensitive element, the k-th sensitivity coefficient is defined as

$$S_k^{ij} \overset{\Delta}{=} \frac{\partial w_{ij}}{\partial x_k} \cdot \frac{x_k}{w_{ij}} \tag{24}$$

From Eq. (23), we therefore obtain the expression

$$S_k^{ij} = \frac{w_{ik}\ w_{kj}}{x_k\ w_{ij}} \tag{25}$$

For computational purposes, it is convenient to express the sensitivity coefficient in terms of the determinant and co-factors of the matrix $\underline{W}^{-1}$. Let

$$\Delta = \det\ \underline{W}^{-1} = \det(\underline{X}^{-1} - \underline{C}) \tag{26}$$

and let Δ_{ij} = co-factor of Δ corresponding to the i-th row, j-th column element. Then

$$w_{ij} = \frac{\Delta_{ji}}{\Delta} . \tag{27}$$

Hence from Eq. (25), we obtain

$$S^{ij}_k = \frac{\Delta_{ij}\,\Delta_{ki}}{x_k\,\Delta_{ji}\,\Delta} \tag{28}$$

For experimental measurement using models of the system, it is convenient to express the sensitivity coefficient in terms of the elements of both the sensitivity matrices $\underline{S}$ and $\underline{\hat{S}}$. Using Eq. (19) in Eq. (25), we obtain the expression

$$S^{ij}_k = \frac{s_{ik}\,x_k\,\hat{s}_{kj}}{w_{ij}} \tag{29}$$

Given two identical models of the system, therefore, the k-th sensitivity coefficient may be measured by means of the arrangement shown in Fig. 3.

In particular, if k = i,

$$S^{ij}_i = \frac{w_{ii}}{x_i} = S^{ii}_i = s_{ii} = \hat{s}_{ii} \tag{30}$$

and if k = j,

$$S^{ij}_j = \frac{w_{jj}}{x_j} = S^{jj}_j = s_{jj} = \hat{s}_{jj} \tag{31}$$

Thus the diagonal elements of either sensitivity matrix have the physical meanings of the conventional scalar sensitivity.

Similarly, formulas for the conventional second order and higher order sensitivities can be obtained. The second order sensitivity can be defined by

$$S^{ij}_{k\ell} \triangleq \frac{x_k x_\ell}{2w_{ij}} \frac{\partial^2 w_{ij}}{\partial x_k \partial x_\ell} \tag{32}$$

From Eq. (23) it is easily shown that

$$S^{ij}_{k\ell} = \frac{w_{i\ell}\,w_{\ell k}\,w_{kj} + w_{ik}\,w_{k\ell}\,w_{\ell j}}{2x_k x_\ell\,w_{ij}}$$

$$= \frac{1}{2}\left(S^{ik}_{\ell}\,S^{ij}_k + S^{i\ell}_k\,S^{ij}_{\ell}\right), \quad k \neq \ell \tag{33}$$

while

$$S^{ij}_{kk} = \frac{w_{ik}\,w_{kj}}{x_k\,w_{ij}}\left(\frac{w_{kk}}{x_k} - 1\right) = S^{ij}_k\left(S^{kk}_k - 1\right) \tag{34}$$

Thus the higher order sensitivities are expressed in terms of the lower order sensitivities and the scalar transfer functions. Equation (22) can be rewritten in terms of these scalar sensitivities in the form of a fractional change of w_{ij} as follows:

$$\frac{\delta w_{ij}}{w_{ij}} = \sum_{k=1}^{n} S_k^{ij} \frac{\delta x_k}{x_k} + \sum_{k,\ell=1}^{n} S_{k\ell}^{ij} \frac{\delta x_k \, \delta x_\ell}{x_k \, x_\ell} + \dots \qquad (35)$$

This formula, in conjunction with Eqs. (28) - (31), Eqs. (33) - (34), and similar formulas for higher order sensitivity coefficients, which may be derived in the same way as above, furnishes all the information needed for a complete sensitivity analysis of a scalar transfer function.

A convenient way of expressing the first order increment of a scalar transfer function in terms of the increments of the sensitive parameters is to define a sensitivity vector:(9)

$$\underline{s} \triangleq \text{col.} \, (S_1^{ij}, S_2^{ij}, \dots . \, S_n^{ij}) \, . \qquad (36)$$

Then if we define the perturbation vector:

$$\underline{p} \overset{\Delta}{=} \text{col.} \, \left(\frac{\delta x_1}{x_1}, \frac{\delta x_2}{x_2}, \dots \frac{\delta x_n}{x_n} \right) , \qquad (37)$$

we have

$$\frac{\delta w_{ij}}{w_{ij}} = \underline{s}^t \, \underline{p} \, . \qquad (38)$$

FLOW GRAPH REDUCTION

The sensitivity analysis derived in the previous two sections are based on a particular flow graph model as shown in Fig. 1, and in addition the sensitive parameters are represented by a diagonal matrix $\underline{X}$. Although the assumed model represents many linear systems of interest, it is not the most general one. In this section we wish to demonstrate that other general flow graph configurations can be reduced to the desired model.

(a) Consider the general configuration of Fig. 4a, where the sensitive set of the system is represented by a diagonal matrix $\underline{X}'$, while the insensitive set is represented by transfer matrices $\underline{A}$, $\underline{B}$, $\underline{D}$ and $\underline{C}'$. Note that the input $\underline{u}'$ and the output $\underline{z}'$ are connected by a direct forward transmission path with transfer matrix D. Off hand, the flow graph does not belong to our model. However, we can augment the flow graph of Fig. 4a. by adding superfluous unit matrices as shown in Fig. 4b. and obtain a model which is of the assumed form. In the augmented flow graph the input, output and the transfer matrices are:

$$\underline{u} = \begin{bmatrix} \underline{u}_1 \\ \underline{u}_2 \\ \underline{u}_3 \end{bmatrix} = \begin{bmatrix} \underline{u}' \\ \underline{0} \\ \underline{0} \end{bmatrix} \qquad \underline{z} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \\ \underline{z}_3 \end{bmatrix} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \\ \underline{z}' \end{bmatrix}$$

$$\underline{X} = \begin{bmatrix} \underline{1} & \underline{0} & \underline{0} \\ \underline{0} & \underline{X}' & \underline{0} \\ \underline{0} & \underline{0} & \underline{1} \end{bmatrix} \qquad \underline{C} = \begin{bmatrix} \underline{0} & \underline{0} & \underline{0} \\ \underline{A} & \underline{C}' & \underline{0} \\ \underline{D} & \underline{B} & \underline{0} \end{bmatrix} \qquad (39)$$

(b) Next, we consider the situations where the transfer matrix of the sensitive set is not diagonal. This can be illustrated by two simple scalar flow graphs in Fig. 5a., where two sensitive elements are incident to a common node and where two sensitive elements originate from a single node. In both situations $\underline{X}$ are not diagonal. By inserting superfluous unit elements we obtain the modified flow graphs in Fig. 5b. In Fig. 5 the heavy lines represent the elements in the sensitive set, the light lines represent the elements in the insensitive set. It is easily seen that the modified flow graphs in Fig. 5b. are in the form of our assumed model.

SENSITIVITY ANALYSIS OF RLC NETWORKS

In precision passive networks it is often necessary to make an extensive sensitivity analysis when the design is completed and when the circuit is built. We demonstrate that the preceding sensitivity analysis can be applied directly to general RLC networks. The method leads to systematic computations and straightforward experimental measurements for a complete sensitivity study.

Consider a general RLC network with many loops and nodes. We are interested in the sensitivity of a particular transfer function with respect to all elements in the network. From the Kirchhoff laws, we have

$$\underline{B}\;\underline{v} = \underline{e}\,, \qquad \underline{Q}\;\underline{i} = \underline{j} \tag{40}$$

where $\underline{B}$ is the fundamental loop matrix, $\underline{Q}$ is the fundamental cut-set matrix, $\underline{v}$ and $\underline{i}$ are the branch voltage and current vectors, respectively. The sources are represented by $\underline{e}$, the voltage source vector in the fundamental loops, and by $\underline{j}$, the current source vector in the fundamental cut-sets. If we partition the branches into links and tree branches and use subscripts 1 and 2 to designate them, respectively, we obtain

$$\begin{aligned} \underline{B}\;\underline{v} &= \begin{bmatrix} \underline{1} & \underline{F} \end{bmatrix} \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \end{bmatrix} = \underline{v}_1 + \underline{F}\;\underline{v}_2 = \underline{e} \\ \underline{Q}\,\underline{i} &= \begin{bmatrix} -\underline{F}^t & \underline{1} \end{bmatrix} \begin{bmatrix} \underline{i}_1 \\ \underline{i}_2 \end{bmatrix} = -\underline{F}^t\,\underline{i}_1 + \underline{i}_2 = \underline{j} \end{aligned} \tag{41}$$

where F is a submatrix of the fundamental loop matrix and F^t is its transpose. Rewriting Eq. (41), we have

$$\begin{bmatrix} \underline{v}_1 \\ \underline{i}_2 \end{bmatrix} = \begin{bmatrix} \underline{0} & -\underline{F} \\ \underline{F}^t & \underline{0} \end{bmatrix} \begin{bmatrix} \underline{i}_1 \\ \underline{v}_2 \end{bmatrix} + \begin{bmatrix} \underline{e} \\ \underline{j} \end{bmatrix} \tag{42}$$

On the other hand, the branch voltage and current vectors are related by the branch equations by a hybrid matrix, H. Thus

$$\begin{bmatrix} \underline{i}_1 \\ \underline{v}_2 \end{bmatrix} = \underline{H}^{-1} \begin{bmatrix} \underline{v}_1 \\ \underline{i}_2 \end{bmatrix} \tag{43}$$

where H is a diagonal matrix which contains all elements of the sensitive set, and is of the form

$$\underline{H} = \text{diag.}\,[\, p^{-1}\underline{S},\ \underline{R},\ p\underline{L},\ p\underline{C},\ \underline{G},\ p^{-1}\underline{\Gamma}] \qquad (44)$$

In the above p is the complex frequency; $\underline{S}$, $\underline{R}$, $\underline{L}$, $\underline{C}$, $\underline{G}$ and $\underline{\Gamma}$ represent matrices of link elastances, link resistances, link inductances, tree branch capacitances, tree branch conductances and tree branch reciprocal inductances, respectively.

By comparing Eqs. (42), (43) and (1), (2) we have the following identifications:

$$\underline{u} = \begin{bmatrix} \underline{e} \\ \\ \underline{j} \end{bmatrix}, \quad \underline{v} = \underline{0}, \quad \underline{y} = \begin{bmatrix} \underline{v}_1 \\ \\ \underline{i}_2 \end{bmatrix} \text{ and } \underline{z} = \begin{bmatrix} \underline{i}_1 \\ \\ \underline{v}_2 \end{bmatrix} \qquad (45)$$

$$\underline{C} = \begin{bmatrix} \underline{0} & -\underline{F} \\ \\ \underline{F}^t & \underline{0} \end{bmatrix} \text{ and } \underline{X} = \underline{H}^{-1} \qquad (46)$$

Thus for any linear RLC network, we can define a transfer matrix which relates $\begin{bmatrix} \underline{e} \\ \underline{j} \end{bmatrix}$ and $\begin{bmatrix} \underline{i}_1 \\ \underline{v}_2 \end{bmatrix}$, and this transfer matrix describes completely the given network. In network problems a scalar transfer function is usually of interest. It relates a link current or tree branch voltage output to a fundamental loop voltage source or a fundamental cut-set current source input. To obtain a complete sensitivity analysis we only need to use Eq. (35) and compute all scalar sensitivities. If the transfer function is w_{ij}, to compute all first order scalar sensitivities we compute first all w_{ik} and w_{kj} for $k=1, 2, \ldots, n$.

COMPARISON OF OVERALL SENSITIVITIES OF SYSTEMS

In many cases of considerable practical importance, a system designer is required to choose from a number of possible synthesis schemes the one least sensitive to the fluctuation of certain parameters. But even in the simplest case involving the synthesis of a single scalar transfer function, where all possible design schemes consist of the same set of sensitive elements, the problem of finding a criterion for comparing the overall sensitivities of two possible schemes is not yet completely solved. If $\underline{s}_1$ and $\underline{s}_2$ be the sensitivity vectors, as defined in Eq. (36), for the two schemes, it is evident that the first system is at least as good as the second for a given perturbation vector $\underline{p}$ if

$$|\, \underline{s}_1^t\ \underline{p} \,| \le |\, \underline{s}_2^t\ \underline{p} \,| \qquad (47)$$

If this inequality holds for any arbitrary $\underline{p}$, the first design may be said to be less sensitive than the second. It is clear, however, that (47) cannot be valid for arbitrary $\underline{p}$ unless either (i) $\underline{s}_1 = 0$ or (ii) $\underline{s}_1$ and $\underline{s}_2$ are colinear and $\|\underline{s}_1\| \le \|\underline{s}_2\|$, for if neither (i) nor (ii) is true, then by choosing $\underline{p}$ orthogonal to $\underline{s}_2$, one can make $\underline{s}_2^t\underline{p} = 0$, while $\underline{s}_1^t\ \underline{p} \neq 0$. It follows, therefore, that a design can be less sensitive than all other possible designs for arbitrary perturbations only if its corresponding sensitivity coefficients are all zero. In theory, such a completely insensitive transfer function can be realized by introducing unity scalar loop transmissions. However, such a scheme is of little practical value, since the unit positive feedback makes the system unstable due to the otherwise non-dominant natural frequencies. It is therefore impossible to design a minimum-sensitivity system whose deviation from the ideal will be less than that for any other possible design, for arbitrary perturbations. The answer to the problem is hence to be sought in a stochastic sense, in terms of the statistical distributions of the sensitive parameters.

REFERENCES

1. H. W. Bode, Network Analysis and Feedback Amplifier Design, D. Van Nostrand Co., Inc., New York; 1945.
2. S. J. Mason, "Feedback theory-some properties of signal flow graphs," Proc. IRE, Vol. 41, pp. 1144-1156; September 1953.
3. J. G. Truxal, Automatic Feedback Control System Synthesis, McGraw-Hill Book Co., Inc., New York; 1955.
4. I. M. Horowitz, "Synthesis of linear, multivariable feedback control systems," IRE Trans. on Automatic Control, Vol. AC-5, pp. 94-105; June, 1960.
5. I. M. Horowitz, Synthesis of Feedback Systems, Academic Press, New York; 1963.
6. I. W. Sandberg, "On the theory of linear multiloop feedback systems," BSTJ, Vol. 42, pp. 355-382; March, 1963.
7. E. S. Kuh, "Some results in linear multiple loop feedback systems," Proc. Allerton Conf. on Circuit and System Theory, pp. 471-487; November, 1963.
8. J. B. Cruz, Jr., and W. R. Perkins, "A new approach to the sensitivity problem in multivariable feedback system design," IEEE Trans. on Automatic Control, Vol. AC-9, pp. 216-223; July, 1964.
9. A. J. Goldstein, and F. F. Kuo, "Multiparameter Sensitivity," IRE Trans. on Circuit Theory, Vol. CT-8, pp. 177-178; June, 1961.

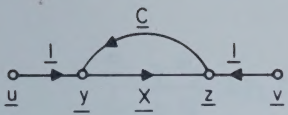
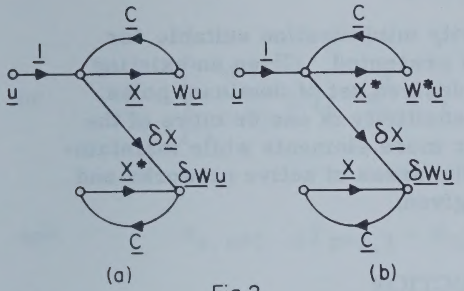
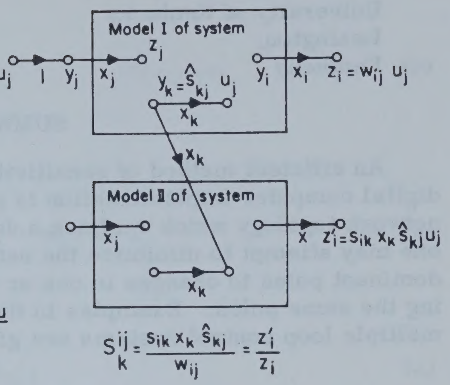


Fig.1.


C
1
u
X
Wu
δX
X*
δWu
C
C
1
u
X*
W*u
δX
X
δWu
C

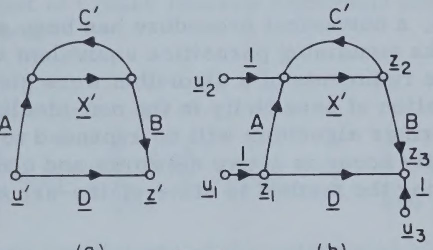
(a) (b)

Fig.2.


Model I of system
u_j
1
y_j
x_j
z_j
$y_k = \hat{s}_{kj}\ u_j$
x_k
y_i
x_i
$z_i = w_{ij}\ u_j$
x_k
Model II of system
x_j
x_k
x
$z_i' = s_{ik}\ x_k\ \hat{s}_{kj}\ u_j$

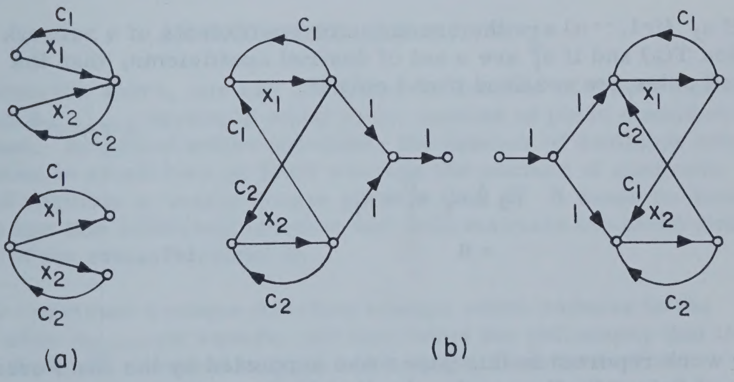
$$S_k^{ij} = \frac{s_{ik}\, x_k\, \hat{s}_{kj}}{w_{ij}} = \frac{z_i'}{z_i}$$

Fig.3.


C′
A
X′
B
u′
D
z′
C′
1
u_2
z_2
A
X′
B
1
z_3
u_1
z_1
D
1
u_3

(a) (b)

Fig.4.


c_1
x_1
x_2
c_2
c_1
x_1
x_2
c_2
c_1
x_1
c_1
1
1
c_2
x_2
1
c_2
c_1
1
x_1
c_2
c_1
x_2
1
c_2

(a) (b)

Fig.5.

A NUMERICAL ALGORITHM FOR THE MINIMIZATION OF SENSITIVITY*

D. A. CALAHAN
University of Kentucky
Lexington,
Kentucky

SUMMARY

An efficient method of sensitivity minimization suitable for digital computer implementation is presented. Given an existing network topology which realizes a desired set of dominant poles, one may attempt to minimize the sensitivity of one or more of the dominant poles to changes in one or more elements while maintaining the same poles. Examples in the areas of active networks and multiple loop control systems are given.

INTRODUCTION

In reference (1), a numerical procedure has been given for synthesizing networks containing parasitics equivalent to idealized networks. Also, the rudiments of a algorithm were given for the subsequent minimization of sensitivity in the non-idealized network. In this paper, the former algorithm will be expanded to include cases which frequently occur in active networks and control systems. The results of applying the method to state-of-the-art sensitivity problems are given.

A REVIEW

If a_i $(i=1, \cdots n)$ are the denominator coefficients of a network function T(s) and if a_i^d are a set of desired coefficients, then the desired poles are realized if and only if

$$\begin{aligned} E_i &\triangleq c_1 \; a_i^d - a_i \\ &= 0 \qquad\qquad i=1 \ldots n \qquad\qquad (1) \end{aligned}$$

---

* The work reported in this paper was supported by the Air Force Office of Scientific Research, Applied Mathematics Division, in part on Grant AF177-63 at the University of Illinois and in part on Grant AF948-65 at the University of Kentucky, Lexington.

Regarding a_i as a function of the elemental admittances y_j, an iteration scheme is defined in the Newton-Raphson sense by letting

$$\sum_{j=1}^{n} \frac{\partial E_i}{\partial y_j} \Delta y_j = - E_i \qquad i=1,\dots r \tag{2}$$

or

$$-\sum_{j=1}^{n} \frac{\partial a_i}{\partial y_j} \Delta y_j + a_i^d \Delta c_1 = a_i - c_1 a_i^d \tag{3}$$

or*

$$A_{r,n+1} \quad \Delta Y_{n+1,1} = B_{r,1} \tag{4}$$

When only a subset of t poles (usually dominant) are to be maintained, Equation 3 becomes

$$-\sum_{j=1}^{n} \frac{\partial a_i}{\partial y_j} \Delta y_j + \sum_{k=o}^{i} a_k^d \Delta c_{i-k} = a_i - \sum_{k=o}^{i} a_k^d c_{i-k} \tag{5}$$

where the c_{i-k} are coefficients of the polynomial containing the non-dominant poles. Hence, there are r-t c_{i-k} terms.

AN EXPANDED ALGORITHM

From the above, one can show that the number of Δy_j's required to make $A_{r,n+r-t}$ square is equal to the number of poles considered dominant. In typical active networks, the number of dominant complex poles is small (two or four) whereas the number of elements allowed variable is nearly always greater than n. It would be desirable to use this additional freedom and still maintain the multi-linear nature of the error functions E_i.

To construct a unique iteration scheme which reduces to the above when $A_{r,n+r-t}$ is square, one may adopt the philosophy that the initial design is assumed to be in some sense close to the final design.

---

*Matrices will be denoted by subscripts giving their orders.

Therefore, one may consider minimizing at each iteration the function

$$\varphi \triangleq \sum_{j=1}^{n} \left(\frac{\Delta y_j}{y_j}\right)^2 \tag{6}$$

subject to the linear constraints imposed by Equation 5.

Defining

$$\Delta x_j = \frac{\Delta y_j}{y_j} \qquad 1 \le j \le n \tag{7}$$

$$= \Delta c_{i-j} \qquad n < j \le m \tag{8}$$

then the problem becomes simply to minimize

$$\varphi = \sum_{j=1}^{n} (\Delta x_j)^2 \tag{9}$$

subject to linear constraints of the form

$$\sum_{j=1}^{m} c_{ij} \Delta x_j = d_i \qquad i = 1, \dots r \tag{10}$$

where $m=n+r-t$. Although this begins to appear as a problem in quadratic programming, the solution is in fact much simpler in concept.

To begin, if the equations represented by Equation 5 are considered as r hyperplanes in m-space, then their intersection is an m-r dimensional hyperplane. If φ contains all m Δx_j, then $\sqrt{\varphi}$ is simply the length of the normal from this hyperplane to the origin in the Δx_j plane. The intersection of this normal is readily found by constructing m-r independent equations describing an m-r dimensional hyperplane which is perpendicular to the former and passes through the origin. The method of generating these equations is straightforward, and is discussed in the Appendix.

In fact, only n of the Δx_j are included in φ; however, the above argument applies after m-n variables have been eliminated.

POSITIVE ELEMENTS

When $A_{r,m}$ is non-square, then if the dominant poles are realizeable at all with the given topology, there are usually a continuum of possible solutions. Having some choice, one may then wish to require all elements to be positive. If φ could be weighted such that multiplying y_i by a constant K were as serious as dividing y_i by K, then y_i could not become negative.

Such weighting can be achieved by defining the Δy_j's from

$$y_j + \Delta y_j = y_j e^{w_j} \tag{11}$$

and minimizing

$$\varphi = \sum_{j=1}^{n} (w_j)^2 \tag{12}$$

Note $w_j=0$ at the beginning of each iteration, so $w_j = \Delta w_j$. Hence, Equation 5 becomes

$$-\sum_{j=1}^{n} y_j \frac{\partial a_i}{\partial y_j} \Delta w_j + \sum_{k=o}^{i} a_k^d \Delta c_{i-k} = a_i - \sum_{k=o}^{i} a_k^d c_{i-k} \tag{13}$$

Equation 5 and 13 are identical except that Δw_j appears in place of Δy_j. But, from Equation 11,

$$y_j + \Delta y_j = y_j + y_j \Delta w_j + y_j \frac{(\Delta w_j)^2}{2} + \dots \tag{14}$$

so that the methods of Equations 5 and 13 approach each other for small Δy_j.

MINIMIZATION OF ROOT SENSITIVITY

Using the usual definition for root sensitivity, i. e.,

$$S_{y_j}^{s_k} \triangleq y_j \frac{\partial s_k}{\partial y_j} \tag{15}$$

where y_j is an admittance
s_k is a root of a polynomial N(s)
$N(s) = N_o(s)(s-s_k)$

then for simple roots the sensitivity is given as

$$S_{y_j}^{s_k} = \frac{\partial N(s)}{\partial y_k} \times \frac{y_k}{N_o(s)} \Bigg|_{s=s_k} \tag{16}$$

For simplicity, Re $S_{y_j}^{s_k}$ is first examined.

Four polynomials required in the formulation of the sensitivity minimization problem are first defined:

1. Since N(s) is linear in y_k, define

$$N(s) = N_1(s) + y_k N_2(s) \tag{17}$$

and

$$N_2(s_k) = \alpha_2 + j\beta_2 \tag{18}$$

2. From the previous definition of the dominant roots

$$N_3(s) = \sum_{i=o}^{m} a_i^d s^i \qquad a_m^d = 1 \tag{19}$$

$$N_4(s) = \sum_{i=o}^{m} c_i s^i \tag{20}$$

and

$$N_4(s_k) = \alpha_4 + j\beta_4 \tag{21}$$

3. Since s_k must be contained in $N_3(s)$, define

$$N_3(s) = N_5(s)(s - s_k) \tag{22}$$

and

$$N_5(s_k) = \alpha_5 + j\beta_5 \tag{23}$$

Then

$$S_{y_j}^{s_k} = \frac{y_j N_2(s)}{N_o(s)} \Bigg|_{s=s_k} \tag{24}$$

$$= \frac{y_j N_2(s)}{N_4(s) N_5(s)} \Bigg|_{s=s_k} \tag{25}$$

$$= \frac{y_j(\alpha_2 + j\beta_2)}{(\alpha_4 + j\beta_4)(\alpha_5 + j\beta_5)} \tag{26}$$

so that

$$\text{Re } S_{y_j}^{s_k} = \frac{y_p\,(\alpha_2\,\alpha_4\,\alpha_5 + \beta_2\,\beta_4\,\alpha_5 - \beta_2\,\alpha_4\,\beta_5 + \alpha_2\,\beta_4\,\beta_5)}{(\alpha_4^2 + \beta_4^2)\,(\alpha_5^2 + \beta_5^2)} \qquad (27)$$

The α_2 and β_2 are readily calculated, since the coefficients must be known as functions of the elements in order to form the partial derivatives of Equation 13.

If $\bar{\alpha}$ represents a desired real part of root sensitivity, E_i is defined as simply ($\text{Re } S_{y_k}^{s_k} - \bar{\alpha}$). Now α_5 and β_5 can be regarded as constants throughout the iteration since these represent the dominant poles which both the initial and final design are assumed to possess. On the other hand, α_2 and β_2 are multi-linear in the y_j's, whereas α_4 and β_4 are linear in the c_j's. Thus, E_i, although not strictly multi-linear in the Δx_i's (Equation 27), will be quadratic only in so far as the non-dominant roots - and hence the c_j's - change as the iteration proceeds.

Similar formulae can be developed for minimization of the imaginary part root sensitivity, of obvious interest for improving frequency stability of oscillators and tuned amplifiers.

EXAMPLES

Example 1.

Let it be desired to design a 10% active filter to operate at 1.59 Mc. using the forward gain block and feedback network given in Figure 1 (frequency scaled by 10^7 and impedance scaled by 10^3). The filter is to have its center frequency and bandwidth separately controlled by two resistances R_1 and R_2 respectively, which are to be used as alignment elements.

Using the method described in ref. (1), the network of Figure 1 is designed so that a dominant pole is present at $s = -.05 \pm j1$. The real part of the root sensitivity with respect to R_1 is to be forced to zero and the imaginary part with respect to R_2 is to be zero. After two minutes of computing time, the final design is given in Figure 1 and the final element values are given in parentheses in Figure 1.

Example 2.

Figure 2 shows a two-loop feedback control system with a dominant pole at $s = -1. \pm j1$. The sensitivity of this dominant root with respect to K is $-.86 \mp j4.15$. It is desired to minimize the magnitude of this sensitivity while keeping the dominant root in place.

A network realization is first made of the control system (Figure 3). In this case the compensating networks are themselves used in the realization, although in general this may impose too great a restriction on the blocks realized (particularly if all network elements are required to remain positive). The minimization was halted after three minutes of computing time, when it became reasonably clear the sensitivity could be reduced without limit by increasing the loop gain. The final sensitivity was $.199 \mp j.056$; the values are given in parentheses in Figure 3.

APPENDIX I

Examples of solution when φ contains all variables

Problem 1: Minimize

$$\varphi_1 = x_1^2 + x_2^2 + x_3^2 \tag{I-1}$$

subject to

$$4\,x_1 + 2\,x_2 + 3\,x_3 = 1 \tag{I-2}$$

Solution: Equation I-1 describes a three-dimensional hyperplane P_1. Therefore we must generate two more independent equations which describe a line passing through the origin and normal to P_1. We write the normal equations of this line as

$$a_2\,x_1 + x_2 = 0 \tag{I-3}$$

$$a_3\,x_1 + x_3 = 0 \tag{I-4}$$

which are certainly independent. The line will be normal to P_1 if and only if

$$4\,a_2 + 2 = 0 \tag{I-5}$$

$$4\,a_3 + 3 = 0 \tag{I-6}$$

yielding $a_2 = -1/2$, $a_3 = -3/4$. Equations I-2, I-3, and I-4 then are solved to yield $x_1 = 4/29$, $x_2 = 2/29$, $x_3 = 3/29$.

Example of Solution When φ Does Not Contain All Variables

Problem 2: Minimize

$$\varphi_2 = x_1^2 + x_3^2 \tag{I-7}$$

subject to

$$4\,x_1 + 2\,x_2 + 3\,x_3 = 1 \qquad \text{(I-8)}$$

$$.5\,x_1 - x_2 = 0 \qquad \text{(I-9)}$$

Solution: Eliminating x_2 in Equations (I-8) and (I-9) yields

$$5\,x_1 + 3\,x_3 = 1 \qquad \text{(I-10)}$$

Equations (I-7) and (I-10) are then solved as in Problem 1, resulting in $x_1 = .147$, $x_2 = .0735$, $x_3 = .0882$

Comment: φ_2 is smaller for the solution of Problem 2 than those of Problem 1^2, as anticipated.

General formulation

Problem 3: Minimize

$$\varphi_3 = \sum_{j=1}^{n} (x_j)^2 \qquad \text{(I-11)}$$

subject to the linearly independent, consistent constraints

$$\sum_{j=1}^{m} c_{ij}\,x_j = d_i \qquad i = 1, 2, \ldots r \qquad \text{(I-12)}$$

or

$$C_{r,m}\,X_{m,1} = D_{r,1} \qquad \text{(I-13)}$$

where*

$$m > r \qquad \text{(I-14)}$$
$$n > m-r \qquad \text{(I-15)}$$

Solution: If $C_{r,m}$ and $X_{m,1}$ are partitioned as

$$C_{r,m} = \begin{bmatrix} C_{r,m-r} & C_{r,r} \end{bmatrix} \qquad \text{(I-16)}$$

$$X_{m,1} = \begin{bmatrix} X_{m-r,1} \\ X_{r,1} \end{bmatrix} \qquad \text{(I-17)}$$

then

$$X_{r,1} = C^{-1}_{r,r}\left(D_{r,1} - C_{r,m-r}\,X_{m-r,1}\right) \qquad \text{(I-18)}$$

$$= G_{r,1} + H_{r,m-r}\,X_{m-r,1} \qquad \text{(I-19)}$$

---

* If $n \leq m-r$, choose all $x_i = 0$ $(i=1, \ldots n)$; then $\varphi_{min} = 0$.

or

$$\begin{bmatrix} H_{r,\,m-r} & -I_{r,\,r} \end{bmatrix} \begin{bmatrix} X_{m-r,\,1} \\ X_{r,\,1} \end{bmatrix} = G_{r,\,1} \tag{I-20}$$

If we now partition $X_{m,\,1}$ as

$$X_{m,\,1} = \begin{bmatrix} X_{n,\,1} \\ X_{m-n,\,1} \end{bmatrix} \tag{I-21}$$

and $G_{r,\,1}$ and $H_{r,\,m-r}$ into

$$G_{r,\,1} = \begin{bmatrix} G_{r-m+n,\,1} \\ G_{m-n,\,1} \end{bmatrix} \qquad H_{r,\,m-r} = \begin{bmatrix} H_{r-m+n,\,m-r} \\ H_{m-n,\,m-r} \end{bmatrix} \tag{I-22}$$

(which is permitted by Equation I-15) so that

$$\begin{bmatrix} H_{r-m+n,\,m-r} & I_{n-m+r,\,n-m+r} \end{bmatrix} X_{n,\,1} = G_{r-m+n,\,1} \tag{I-23}$$

Thus, we have a matrix equation in the same m unknown x_i's contained in φ_3.

The equations describing the normal hyperplane are now appended by forming the augmented matrix

$$\begin{bmatrix} H_{r-m+n,\,m-r} & -I_{n-m+r,\,n-m+r} \\ J_{m-r,\,r-m+n} & I_{m-r,\,m-r} \end{bmatrix} X_{m,\,1} = \begin{bmatrix} G_{r-m+n,\,1} \\ O_{m-r,\,1} \end{bmatrix}$$

where $J_{m-r,\,r-m+n}$ is a matrix of unknowns.

(I-24)

Since each new hyperplane represented by each appended equation must be perpendicular to all the r-m+n constraint hyperplanes, r-m+n equations similar to Equation I-5 and I-6 can be generated in r-m+n unknowns (the number of columns of $J_{m-r,\,r-m+n}$). Thus, $J_{m-r,\,r-m+n}$ can be determined and the x_i found for i=1, 2, . . . n.

It is clear the above is a purely formal derivation, since the existence of matrix inverses is always assumed where necessary to continue the argument.

REFERENCES

1. Calahan, D. A. , "Computer Solution of the Networks Realization Problem," Proc. Second Allerton Conference, University of Illinois, Urbana, Illinois; 1964

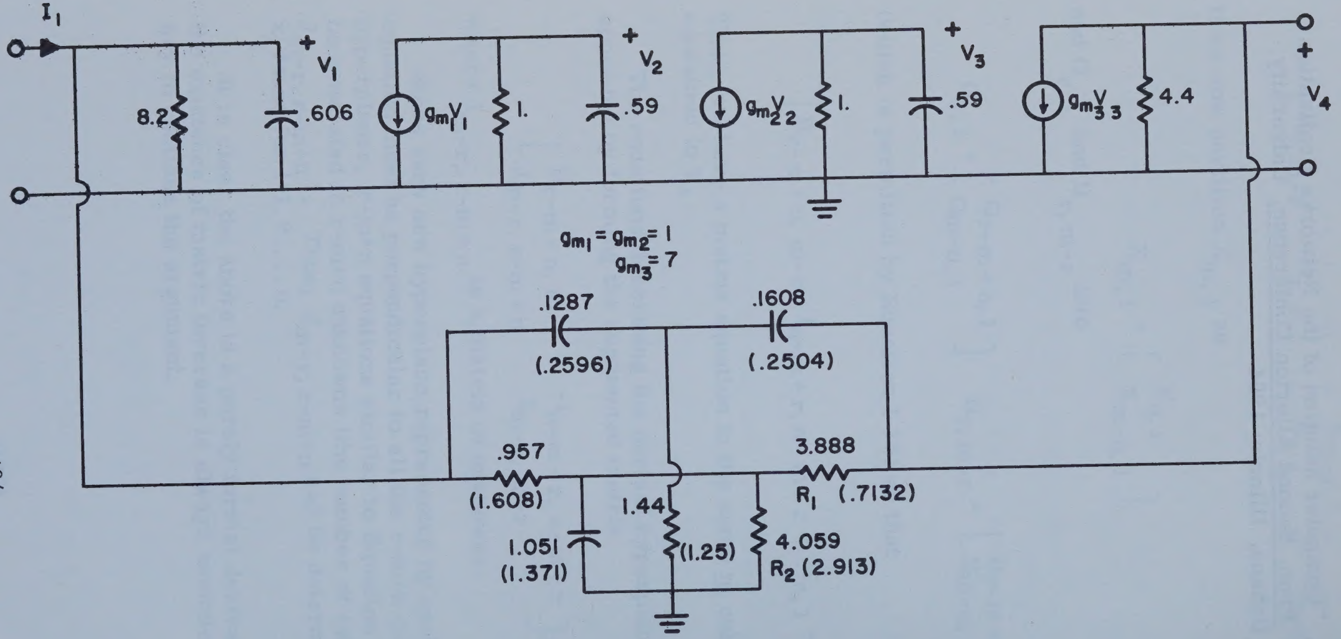

I_1
+ V_1
8.2
.606
$g_{m_1}V_1$
1.
+ V_2
.59
$g_{m_2}V_2$
1.
+ V_3
.59
$g_{m_3}V_3$
4.4
+ V_4
$g_{m_1} = g_{m_2} = 1$
$g_{m_3} = 7$
.1287
(.2596)
.1608
(.2504)
.957
(1.608)
1.44
(1.25)
1.051
(1.371)
3.888
R_1 (.7132)
4.059
R_2 (2.913)

Figure 1. Active RC bandpass filter.

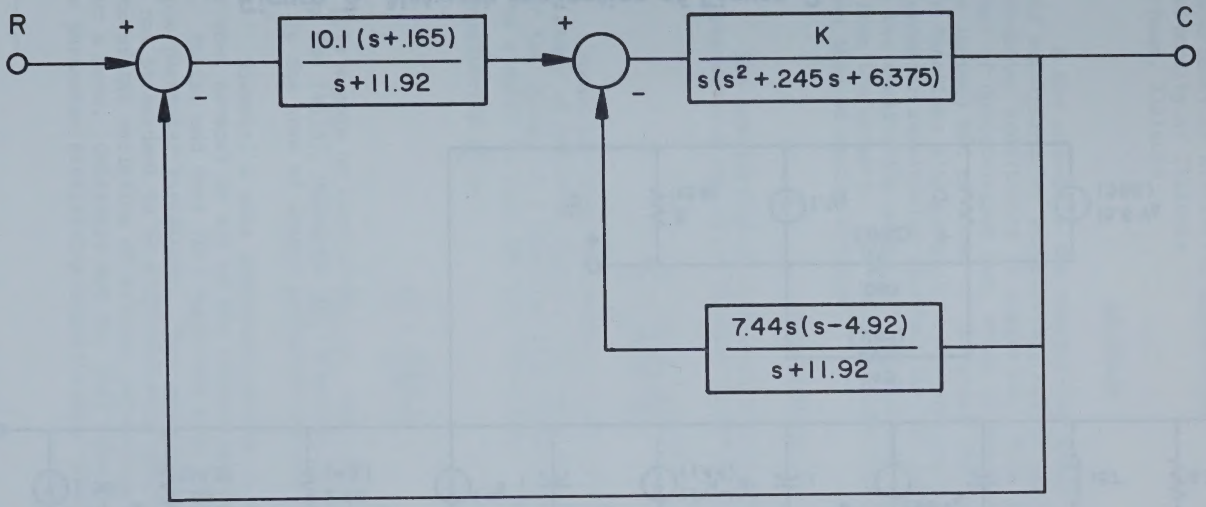

R
+
-
10.1 (s+.165) / s+11.92
+
-
K / s(s² +.245 s + 6.375)
C
7.44s(s-4.92) / s+11.92

Figure 2 Block diagram of example (K=1.98).

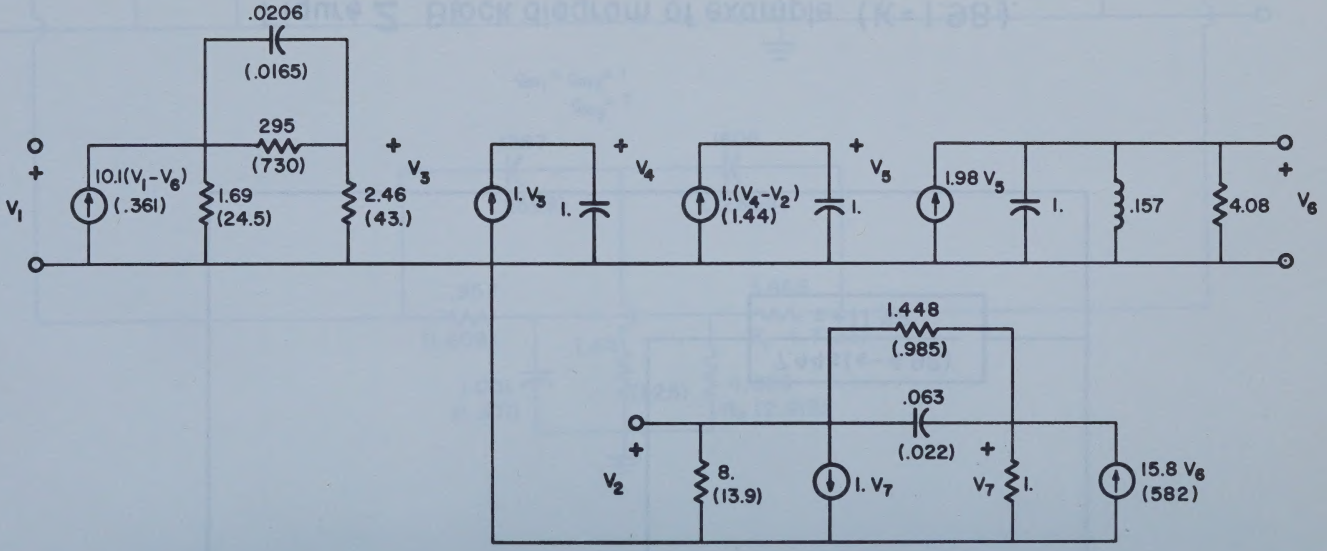

.0206
(.0165)
295
(730)
+
V_1
$10.1(V_1-V_6)$
(.361)
1.69
(24.5)
2.46
(43.)
+
V_3
$1. V_3$
1.
+
V_4
$1.(V_4-V_2)$
(1.44)
1.
+
V_5
$1.98\ V_5$
1.
.157
4.08
+
V_6
1.448
(.985)
.063
(.022)
+
V_2
8.
(13.9)
$1. V_7$
+
V_7
1.
$15.8\ V_6$
(582)

Figure 3. Network realization of Figure 2.

ON MULTIPARAMETER SENSITIVITY*

SAMUEL C. LEE
Department of Electrical Engineering
University of Illinois
Urbana, Illinois

ABSTRACT

A new concept - the sensitivity group - is introduced to the study of multiparameter sensitivity, and an improved definition of multiparameter sensitivity is suggested. This definition, more general than that of Goldstein and Kuo, makes it possible to compare two (linear lumped) RLC networks of different topology realizing the same network functions with respect to multiparameter sensitivity. Some multiparameter sensitivity theorems for RLC networks are given, which permit the sensitivity groups of any RLC network to be directly determined from the network. The multiparameter sensitivity of any RLC network can thus be easily obtained.

I. INTRODUCTION

The sensitivity of a network function T which depends upon a single parameter x was first defined by Bode [1]. His definition is

$$S_x^T = \left[\frac{d(\ell n\ T)}{d(\ell n\ x)}\right]^{-1} \tag{1}$$

Mason [2] took the reciprocal of (1) to define sensitivity in the study of feedback theory. Horowitz and Truxal [3] extended the idea of the change of a function with respect to parameter change to apply to a shift in pole and zero locations. They defined the pole-zero sensitivity S_x^p as

$$S_x^p = \frac{d(\ell n\ p)}{d(\ell n\ x)} \tag{2}$$

where p is a pole or zero of T(x,s). Since then, much work has been done by Kuo [4][5], Mikulski [6], Horowitz [7], Calahan [8], Ur [9] and others in the area of single parameter sensitivity.

However, a more important problem is to consider the sensitivity when several or all of the network parameters change simultaneously. Hakimi and Cruz [10] suggested some sensitivity measures with multiple parameter variations. They postulated that the tolerance could not be assumed to be differentially small. Fu [11] studied the effects of the variation of several parameters on the location of a pole or a zero. Goldstein and Kuo [12] extended Mason's definition of single parameter sensitivity significantly to the multiparameter case.

---

*This work was supported by the Joint Services Electronics Programs (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract No. DA 28 043 AMC 00073(E).

In this paper, the treatment of multiparameter sensitivity problem is based on the concept of sensitivity group. A new definition of multiparameter sensitivity is obtained, which is more general than that of Goldstein and Kuo. Sensitivity groups in RLC networks are studied.

II. BACKGROUND

Let $T(x_1,x_2,\ldots,x_n,s)$ be a network function of n parameters $x_1,x_2,\ldots,x_n$, where s, the complex frequency, is considered as a fixed variable. The fractional change of T for small variations is (neglecting the second-order effects)

$$\frac{\Delta T}{T} = d(\ell n\ T) = \sum_i \frac{\partial(\ell n\ T)}{\partial(\ell n\ x_i)}\, d(\ell n\ x_i) \tag{3}$$

Goldstein and Kuo [12] considered the set of fractional parameter increments as a vector $\overrightarrow{dY} = \left(d(\ell n\ x_i),\ldots,d(\ell n\ x_n)\right)$. Letting $y_i = \ell n\ x_i$, (3) can be expressed as the scalar product.

$$\frac{\Delta T}{T} = \vec{\nabla}\,[\ell n\ T(y_1,\ldots,y_n)]\cdot\overrightarrow{dY} \tag{4}$$

The multiparameter sensitivity $\vec{S}^T$ is

$$\vec{S}^T \equiv \vec{\nabla}\,[\ell n\ T(y_1,\ldots,y_n)]. \tag{5}$$

The magnitude of the multiparameter sensitivity function,

$$|\vec{S}^T| = (\vec{S}^T\cdot\vec{S}^T)^{\frac{1}{2}} \tag{6}$$

gives the maximum rate of change of ℓn T with respect to $(\ell n\ x_i)$.

III. SENSITIVITY GROUP

Goldstein and Kuo's definition assumes that the n parameters of T can vary simultaneously and independently and thus can be considered as n linearly independent vectors. In this paper, a new suggestion is made. We would like to take both the type of network function T and the nature of the variations of T due to the parameters into consideration.

Let T be a network function* of n parameters, $x_1,x_2,\ldots,x_n$ and $X = \{x_1,x_2\ldots x_n\}$ be the set of parameters. Suppose a parameter $x_i \in X$ has varied an arbitrary amount Δx_i and the variation of T due to Δx_i is $\Delta T|_{\Delta x_i}$. If there exists a parameter $x_j \in X$ such that $\Delta T|_{\Delta x_i} = \Delta T|_{\Delta x_j}$ and $\Delta x_j = k_x^{(ji)} \Delta x_i$ for all s ($k_x^{(ji)}$ is a non-zero constant independent of s), then any variation of T due to Δx_i can equivalently be obtained by Δx_j, where Δx_i, Δx_j are linearly proportional (Figure 1). Two parameters x_i, x_j of T having the property stated above, are said to be linearly dependent with respect to T. i.e., any variation of T due to one of the two parameters can be realized by the other for all s. Taking the scaling effect of T into account, in the broad sense, we make the following definitions:

---

*In this paper we exclude voltage ratio function and current ratio function.

Definition 1 (Linear dependence). Let $D = \{x_1, x_2, \ldots, x_k\}$ be a set of parameters of X. The parameters $x_1, x_2, \ldots, x_k$ of T are linearly dependent with respect to T if x_i, $x_j \in D$, and for any variation Δx_i, there exist Δx_j, $k_T^{(ji)}$ and $k_x^{(ji)}$ such that

$$\Delta T|_{\Delta x_j} = k_T^{(ji)} \, \Delta T|_{\Delta x_i} \quad \text{and} \quad \Delta x_j = k_x^{(ji)} \Delta x_i \tag{7}$$

$$i, j = 1, 2, \ldots, k$$

where $k_T^{(ji)}$, $k_x^{(ji)}$ are non-zero constants, not functions of s.

Definition 2 (Linear independence). The parameters $x_1, x_2, \ldots, x_k$ of a network function T are linearly independent with respect to T if they are not linearly dependent.

Theorem 1 Let T be a network function of n parameters $X = \{x_1, x_2, \ldots, x_n\}$. Define a relation R on the set X as follows: "$x_i R x_j$ if and only if x_i, $x_j \in X$ and the relation of (7) holds for x_i and x_j". The R is an equivalence relation on X.

Proof: See Reference [14].

Corollary 1.1 R is an equivalence relation on every subset of X.

Now we are in the position to define a sensitivity group.

Definition 3 (Sensitivity group). A set $\mathcal{S}$ of parameters of X is called a sensitivity group with respect to T if $\mathcal{S}$ satisfies the following two conditions:

(1) All elements of $\mathcal{S}$ are linearly dependent with respect to T.

(2) Let x_i be an element in $\mathcal{S}$. $\mathcal{S}$ is the largest R - equivalence class determined by x_i.

Our next task is to show that for any network function T of n parameters $x_1, x_2, \ldots, x_n$, the set $X = \{x_1, x_2, \ldots, x_n\}$ can be always partitioned into $m(\leq n)$ disjoint sensitivity groups, or m disjoint R-equivalence classes, $X = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m\}$. We shall also prove that this partition is unique with respect to T. Before we can proceed, we need the following lemma:

Lemma Let T be a network function of n parameters $x_1, x_2, \ldots, x_n$ and $\mathcal{S}_1$, $\mathcal{S}_2$, be two sensitivity groups. Then either $\mathcal{S}_1 \cap \mathcal{S}_2 = \phi$ or $\mathcal{S}_1 = \mathcal{S}_2$.

Proof: Suppose $\mathcal{S}_1 \cap \mathcal{S}_2 \neq \phi$. Let $x_i \in \mathcal{S}_1 \cap \mathcal{S}_2$, i.e. $x_i \in \mathcal{S}_1$ and $x_i \in \mathcal{S}_2$. By definition of sensitivity group, $\mathcal{S}_1$, $\mathcal{S}_2$ are R-equivalence classes determined by x_i. Moreover, $x_i \in \mathcal{S}_1$, $\mathcal{S}_1$ is the largest R-equivalence class determined by x_i, thus $\mathcal{S}_2 \subseteq \mathcal{S}_1$. Similarly, since $x_i \in \mathcal{S}_2$, we have $\mathcal{S}_1 \subseteq \mathcal{S}_2$. Hence $\mathcal{S}_1 = \mathcal{S}_2$.

Using the lemma, we can prove the following theorem [14].

Theorem 2 Let T be a network function of n parameters $X = \{x_1, x_2, \ldots, x_n\}$. X can always be partitioned into a set of disjoint sensitivity groups, and such a partition is unique.

IV. A NEW DEFINITION OF MULTIPARAMETER SENSITIVITY

Recall (3)

$$\frac{\Delta T}{T} = d(\ell n\ T) = \sum_{i=1}^{n} \frac{\partial(\ell n\ T)}{\partial(\ell n\ x_i)} d(\ell n\ x_i)$$

where the variations of x_i's are assumed to be very small.

Suppose the sensitivity partition of $X = \{x_1, x_2, \ldots, x_n\}$ is

$$X = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m\}$$

where $\mathcal{S}_1 = \left\{x_1, \ldots, x_{n_1}\right\}, \mathcal{S}_2 = \left\{x_{n_1+1}, \ldots, x_{n_2}\right\}, \ldots, \mathcal{S}_m = \left\{x_{n_{m-1}+1}, \ldots, x_n\right\}$

as shown in Figure 2. (3) can be written as

$$\frac{\Delta T}{T} = d(\ell n\ T) = \sum_{i_1=1}^{n_1} \frac{\partial(\ell n\ T)}{\partial\left(\ell n\ x_{i_1}\right)} d\left(\ell n\ x_{i_1}\right) + \sum_{i_2=n_1+1}^{n_2} \frac{\partial(\ell n\ T)}{\partial\left(\ell n\ x_{i_2}\right)} d\left(\ell n\ x_{i_2}\right)$$

$$+ \ldots + \sum_{i_m=n_{m-1}+1}^{n} \frac{\partial(\ell n\ T)}{\partial\left(\ell n\ x_{i_m}\right)} d\left(\ell n\ x_{i_m}\right) \qquad (8)$$

Define the sensitivity-group increments by

$$d\,\ell n\, s_j = \frac{\displaystyle\sum_{i_j=n_{j-1}+1}^{n_j} \frac{\partial(\ell n\ T)}{\partial\left(\ell n\ x_{i_j}\right)}\, d\left(\ell n\ x_{i_j}\right)}{\displaystyle\sum_{i_j=n_{j-1}+1}^{n_j} \frac{\partial(\ell n\ T)}{\partial\left(\ell n\ x_{i_j}\right)}}, \qquad j = 1,2,\ldots,m \quad (9)$$

where $n_0 = 0$ and $n_m = n$.

Letting

$$S_j = \sum_{i_j=n_{j-1}+1}^{n_j} \frac{\partial(\ell n\ T)}{\partial\left(\ell n\ x_{i_j}\right)}, \qquad j = 1,2,\ldots,m \quad (10)$$

(3) can be expressed as

$$\frac{\Delta T}{T} = S_1\ d\ \ell n\ s_1 + S_2\ d\ \ell n\ s_2 +,\ldots,+ S_m\ d\ \ell n\ s_m \quad (11)$$

Based on the concept of sensitivity group, we can consider the sensitivity group increments as a vector.

$$\vec{ds} = d\ \ell n\ \mathcal{S}_1\ \vec{s}_1 + d\ \ell n\ \mathcal{S}_2\ \vec{s}_2 + \ldots + d\ \ell n\ \mathcal{S}_m\ \vec{s}_m \quad (12)$$

where $\vec{s}_i$'s are unit vectors. (11) becomes the scalar product

$$\frac{\Delta T}{T} = \left[S_1\ \vec{s}_1 + S_2\ \vec{s}_2 + \ldots + S_m\ \vec{s}_m\right] \cdot \vec{ds} \quad (13)$$

Therefore, we suggest that multiparameter sensitivity S^T be defined as

$$\vec{S}^T \equiv S_1\ \vec{s}_1 + S_2\ \vec{s}_2 + \ldots + S_m\ \vec{s}_m \quad (14)$$

where

$$S_j = \sum_{i_j=n_{j-1}+1}^{n_j} \frac{\partial(\ell n\ T)}{\partial\left(\ell n\ x_{i_j}\right)} \qquad j = 1,2,\ldots,m$$

= the sum of the single parameter sensitivity of all elements of the j^{th} group

$$= \sum_{i_j=n_{j-1}+1}^{n_j} S^T_{x_{i_j}}$$

This definition of multiparameter sensitivity includes Goldstein and Kuo's definition as a special case when each sensitivity group has only one parameter. From Theorem 2, we find that the expression of $\vec{S}^T$ in (14) is unique.

V. SOME THEOREMS ON THE MULTIPARAMETER SENSITIVITY FOR RLC NETWORKS

When we want to find the multiparameter sensitivity $\vec{S}^T$ for a specific network function T, we first need to find the sensitivity partition of the set X. It was pointed out that the normal procedure of determining the sensitivity partition is tedious; we need to examine all pairs of parameters and see whether they are in one group. However, in most practical situations this testing is not necessary, e.g., in determining the sensitivity partition for a network function of an (linear lumped) RLC network. The network function of an RLC network has a special feature, that is, each element of the network is a parameter of T. The following part of this paper is devoted to a discussion of the multiparameter sensitivity of RLC networks. This will give an insight into the nature of the sensitivity groups.

Definition 4 (Element Group). Let N be an RLC network of n elements $e_1, e_2, \ldots, e_n$, T be a network function of N and $X = \{e_1, e_2, \ldots, e_n\}$. Let $E = \{e_1, e_2, \ldots, e_k\} \equiv X$ be a set of elements of an RLC network N, and N_E be the subnetwork containing the elements $e_1, e_2, \ldots, e_k$. E is said to be an element group with respect to T if it satisfies the following:

(1) all e_i's are of the same element kind.

(2) N_E can be replaced by an equivalent element of the same element kind as e_i's without changing the rest of the element values and the configuration of N.

(3) there does not exist element e_{k+1} of N other than e_i's in E such that $e_1, e_2, \ldots, e_k$, e_{k+1} have the above properties.

Definition 5 (Series Subnetwork Group and Parallel Subnetwork Group). Let T be a network function of an RLC network N. Suppose N can be decomposed into $N_1, N_2, \ldots, N_n, N'$ subnetworks where N_i's and N' are grounded networks. Let $T_1, T_2, \ldots, T_n, T'$ be the network functions of $N_1, N_2, \ldots, N_n, N'$, respectively, where T_i's and T' are of the same type as T. The set of subnetworks $N_1, N_2, \ldots, N_n$ are said to be a series subnetwork group (or a parallel subnetwork group) with respect to T if

(1) N_i's are connected in series (or in parallel)

(2) $T = T_1 + T_2 + \dots + T_n + T'$

$T_i = a_{ij}\,T_j$ for $i,j = 1,2,\dots,n$. Moreover, any N_i can be obtained from N_j by multiplying a magnitude scaling factor.

(3) there does not exist subnetwork N_{n+1} in N' such that $N_1, N_2, \dots, N_n, N_{n+1}$ have above properties.

Definition 6 (Simple and Multiple RLC Network). Let T be a network function of an RLC network N. N is called a *multiple RLC network with respect to T* if N contains at least one series or parallel subnetwork group. If $\overline{N}$ is not multiple, we say N is *simple with respect to T*.

Theorem 3 Let T be a network function of an RLC network N, and N be simple with respect to T. A set of elements of N forms a sensitivity group *if and only if* it is an element group.

Theorem 4 Let T be a network function of an RLC network. Suppose N is multiple with respect to T. Let $N_1, N_2, \dots, N_n$ be a series or parallel subnetwork group with respect to T. Let $e_i^{(N_1)}, e_i^{(N_2)}, \dots, e_i^{(N_n)}$ be the corresponding elements of each subnetwork $N_1, N_2, \dots, N_n$ (refer to Figure 3). $e_i^{(N_1)}, e_i^{(N_2)}, \dots, e_i^{(N_n)}$ form a sensitivity group.

The proofs of Theorem 3 and 4 are rather detailed, and are given in Reference [14].

Corollary 4.1 If each corresponding element $e_i^{(N_j)}$, $j = 1,2,\dots,n$, of N_j is replaced by an element group $E_i^{(N_j)}$, then $E_i^{(N_1)}, E_i^{(N_2)}, \dots, E_i^{(N_n)}$ form a sensitivity group.

Now attention is called to the case as shown in Figure 4, where e's can either be R, L, or C and e's and e''s are of the same element kind. It can be shown that e'_{s_i}, e''_{s_i} form a sensitivity group.

From this example, it reveals that the sensitivity groups of a network is also a function of the locations of the ports. For example, consider the network N_a in Figure 5(a) which gives fourth-order Butterworth low-pass response with respect to ports 11' and 22' and fourth-order Butterworth high-pass response with respect to ports 33' and 44'. The multiparameter sensitivity vectors with respect to ports 11'-22' and ports 33'-44' are

$$\begin{aligned} \vec{S}^{T}_{11'-22'} &= \left(\sum_{i=1}^{2} \frac{\partial\, \ell n\, T}{\partial\, \ell n\, L_i}\right) \vec{s}_{L_{1,2}} + \frac{\partial\, \ell n\, T}{\partial\, \ell n\, L_3}\, \vec{s}_{L_3} + \sum_{i=1}^{2} \frac{\partial\, \ell n\, T}{\partial\, \ell n\, C_i}\, \vec{s}_{C_i} \\ \vec{S}^{T}_{33'-44'} &= \left(\sum_{i=1}^{2} \frac{\partial\, \ell n\, T}{\partial\, \ell n\, C_i}\right) \vec{s}_{C_{1,2}} + \frac{\partial\, \ell n\, T}{\partial\, \ell n\, C_3}\, \vec{s}_{C_3} + \sum_{i=1}^{2} \frac{\partial\, \ell n\, T}{\partial\, \ell n\, L_i}\, \vec{s}_{L_i} \end{aligned} \qquad (15)$$

The network N_b in Figure 5(b) is obtained from N_a after a frequency transformation which gives Butterworth band-pass response and Butterworth band-elimination response with respect to ports 11'-22' and 33'-44' respectively. $\vec{S}^{T}_{11'-22'}$ of N_b has two element groups whereas $\vec{S}^{T}_{33'-44'}$ has none.

In summary, from the above theorems we can determine the sensitivity groups for a given RLC network (with ports and type of network function specified) by inspection.

VI. SENSITIVITY MEASURES

Suppose that two networks N_1, N_2 of different topology realize the same network function T. How can we decide which structure has less sensitivity? If we use $|\vec{S}^{T}|$ of Goldstein and Kuo's definition as a sensitivity measure, we can not determine which structure is better. Each element of N_1, N_2 can be arbitrarily split into a member of elements of the same element kind without changing the topology of the network and each single network can be successively replaced by a number of networks in series or in parallel such that the topology of N_1, N_2 still remain the same. Therefore we suggest that $|\vec{S}^{T}|$ which was determined on the basis of the definition for multiparameter sensitivity (Section IV, Eq. (14)) should be used as a sensitivity measure. The splitting in elements, or the replacing of single network by a number of networks in series or parallel, as described above, will give the same multiparameter sensitivity vector $\vec{S}^{T}$ as the original one.

For the simplest case, consider the open-circuit transfer function z_{12} of the network N_1, N_2, N_3 in Figure 6. Suppose that

$$R_1 = \frac{m}{n} R_0; \quad R_2 = \frac{n-m}{n} R_0; \quad R_3 = \frac{n}{m} R_0; \quad R_4 = \frac{nm}{m(n-m)} R_0 \tag{16}$$

where m, n are positive constant, and n > m. Then

$z_{12}^{(N_1)} = z_{12}^{(N_2)} = z_{12}^{(N_3)} = R_0$. The single parameter sensitivity of each element is

$$S_{R_0}^{z_{12}} = \frac{d\,\ell n\,R_0}{d\,\ell n\,R_0} = 1; \quad S_{R_1}^{z_{12}} = \frac{R_1}{R_1 + R_2}; \quad S_{R_2}^{z_{12}} = \frac{R_2}{R_1 + R_2}; \quad S_{R_3}^{z_{12}} = \frac{R_4}{R_3 + R_4};$$

$$S_{R_4}^{z_{12}} = \frac{R_3}{R_3 + R_4} \tag{17}$$

The multiparameter sensitivity of N_1, N_2, N_3 are

$$\vec{S}_{N_1}^{z_{12}} = 1\cdot\vec{r}_0$$

$$\vec{S}_{N_2}^{z_{12}} = \left(\frac{R_1}{R_1 + R_2} + \frac{R_2}{R_1 + R_2}\right)\vec{r}_{12} = 1\cdot\vec{r}_{12} \tag{18}$$

$$\vec{S}_{N_3}^{z_{12}} = \left(\frac{R_4}{R_3 + R_4} + \frac{R_3}{R_3 + R_4}\right)\vec{r}_{34} = 1\cdot\vec{r}_{34}$$

and

$$\left|\vec{S}_{N_1}^{z_{12}}\right| = \left|\vec{S}_{N_2}^{z_{12}}\right| = \left|\vec{S}_{N_3}^{z_{12}}\right| = 1 \tag{19}$$

From the physical point of view, the result of (19) is reasonable: the maximum rate of change of ($\ell n\ z_{12}$) with respect to ($\ell n\ e_i$) should be invariant for the three networks N_1, N_2, and N_3. For a more general case, consider the transfer function T of the network N_4 of Figure 6. The multiparameter sensitivity of N_4 is

$$\vec{S}_{N_4}^{z_{12}} = \left(\sum_{i=1}^{n} S_{e_i}^{T}\right)\vec{r}$$

It has been shown [13] that

$$\sum_{i=1}^{n} S_{e_i}^{T} = 1$$

where T is an impedance network function.
n is the number of elements.
e_i is a resistance (R), inductance (L) or elastance (1/C)

Thus $\left|\vec{S}_{N_4}^{z_{12}}\right| = 1.$

Next consider the transfer function T of a RLC network N_5 in Figure 7 containing a single-element sensitivity group e. N_6 in Figure 7 is the same network as N_5 by splitting e into an element group. We can show that the splitting of e does not change the maximum rate of change $\vec{S}_{N_5}^{T}$ of N_5 as follow:

$$S_e^T = \frac{\partial\ \ell n\ T}{\partial\ \ell n\ e} = \frac{1}{T}\,\frac{T}{e}\cdot e$$

$$\sum_{i=1}^{n} S_{e_i}^{T} = \sum_{i=1}^{n} \frac{\partial \ell n\ T}{\partial \ell n\ e_i} = \frac{1}{T}\frac{\partial T}{\partial e} \sum_{i=1}^{n} \frac{\partial e}{\partial e_i} e_i$$

$$= \frac{1}{T}\frac{\partial T}{\partial e} \cdot e \cdot \sum_{i=1}^{n} \frac{1}{e}\frac{\partial e}{\partial e_i} e_i = \frac{1}{T}\frac{\partial T}{\partial e} \cdot e \qquad (20)$$

where e is either R, L or 1/C. Note that the expression in the parenthesis can be written as $\sum_{i=1}^{n} \frac{1}{es}\frac{\partial (es)}{\partial e_i} e_i$ if e is an inductance and $\sum_{i=1}^{n} \frac{1}{e/s}\frac{\partial (e/s)}{\partial e_i} e_i$ if e is an elastance.

Since the simple parameter sensitivity of $T(e_i,s)$, $S_{e_i}^{T}(j\omega)$, can be expressed in terms of the rate of change of the attenuation (its real part) and the rate of change of the phase (its imaginary part).

$$S_{e_i}^{T}(j\omega) = \frac{d\alpha(\omega)}{de_i} e_i + j\,\frac{d\beta(\omega)}{de_i} e_i \qquad (21)$$

where $\alpha(\omega) = \ell n\,|\,T(j\omega)\,|$, the attenuation
$\beta(\omega) = \arg T(j\omega)$, the phase

For multiparameter sensitivity, the multiparameter sensitivity of a network function $T = T(e_1, e_2, \ldots, e_n, s)$, $\vec{S}^{T}(j\omega)$ can be written as

$$\vec{S}^{T}(j\omega) = \mathrm{Re}\,[\vec{S}^{T}(j\omega)] + j\ \mathrm{Im}\,[\vec{S}^{T}(j\omega)] \qquad (22)$$

The norm of Re $[\vec{S}^{T}(j\omega)]$ and Im $[\vec{S}^{T}(j\omega)]$ are

$$\left\| \mathrm{Re}\ [\vec{S}^{T}(j\omega)] \right\| = \sqrt{\sum_{i=1}^{m} \Big[\mathrm{Re}\ [S_j(j\omega)] \Big]^2}$$

$$= \sqrt{\sum_{i=1}^{m} \left[\sum_{i_j=n_{j-1}+1}^{n_j} \frac{\partial \alpha(\omega)}{\partial e_{i_j}} e_{i_j} \right]^2} \qquad (23)$$

$$\left\| \operatorname{Im} [\vec{S}^T(j\omega)] \right\| = \sqrt{\sum_{j=1}^{m} \Big[\operatorname{Im} [S_j(j\omega)] \Big]^2}$$

$$= \sqrt{\sum_{j=1}^{m} \left[\sum_{i_j=n_{j-1}+1}^{n_j} \frac{\partial \beta(\omega)}{\partial e_{i_j}} e_{i_j} \right]^2} \qquad (24)$$

where $n_0 = 0$ and $n_m = n$, which give the maximum rate of change of the attenuation with respect to $(\ell n\ e_i)$ and the maximum rate of change of the phase with respect to $(\ell n\ e_i)$. Therefore, $\left\| \operatorname{Re} [\vec{S}^T(j\omega)] \right\|$ and $\left\| \operatorname{Im} [\vec{S}^T(j\omega)] \right\|$ give a sensitivity measure for the attenuation and the phase of the network function.

VII. CONCLUSION

The concept of sensitivity group has been introduced in giving a new definition of multiparameter sensitivity. When each sensitivity group has only one parameter, the new definition agrees with that of Goldstein and Kuo. Some properties of sensitivity group for RLC networks were examined. In particular, the necessary and sufficient conditions for a set of circuit elements of an RLC network to be a sensitivity group have been given, and enable us to determine the sensitivity groups of an RLC network directly from the network without actually through the normal testing procedure. Finally, some multiparameter sensitivity measures were presented to determine the multiparameter sensitivities of two RLC networks of different topology are compared.

ACKNOWLEDGMENT

The author wishes to express his gratitude to Professor M. E. Van Valkenburg for his valuable advice and suggestion in this work. Thanks are also due Professor J. B. Cruz, Jr. for his helpful comments.

REFERENCES

1. Bode, H. W., "Network Analysis and Feedback Amplifier Design", D. Van Nostrand Co., Inc., Princeton, N.J., 1945.

2. Mason, S. J., "Feedback Theory - Some Properties of Signal Flow Graphs", Proc. IRE, Vol. 41, pp. 1144-1156, September 1953.

3. Truxal, J. G. and Horowitz, I. M., "Sensitivity Considerations in Active Network Synthesis", Proc. Second Midwest Symp. on Circuit Theory, Michigan State University, East Lansing, Michigan, p. 6-1 to 6-11, December 1956.

4. Kuo, F. F., "Pole-zero Sensitivity in Network Functions", IRE Trans. on Circuit Theory, Vol. CT-5, No. 4, p. 372, December 1958.

5. Kuo, F. F., "Sensitivity of Transmission Zeros in RC Network Synthesis", IRE Intern'l Convention Record, New York, N.Y., Part 2 pp. 18-22, March 1959.

6. Mikulski, J. J., "A Correlation Between Classical and Pole-zero Sensitivity", Tech. Note No. 5, Contract No. AF49(638)-63, Electrical Engineering Research Laboratory, University of Illinois 1959.

7. Horowitz, I. M., "Optimization of Negative-Impedance Conversion Methods of Active RC Synthesis", IRE Trans. on Circuit Theory, Vol. CT-6, No. 3, pp. 296-303, September 1959.

8. Calahan, D. A., "Sensitivity Minimization in Active RC Synthesis", IRE Trans. on Circuit Theory, Vol. CT-9, No. 1, pp. 38-42, March 1961.

9. Ur, H., "Root Locus Properties and Sensitivity Relations in Control Systems", IRE Trans. on Automatic Control, Vol. AC-5, No. 1, pp. 57-65, January 1960.

10. Hakimi, S. L. and Cruz, J. B., "Measures of Sensitivity for Linear Systems with Large Multiple Parameter Variation", IRE WESCON Convention Record, Part 2, pp. 109-115, August 1960.

11. Fu, K. S., "Sensitivity of a Linear System with Variations of One or Several Parameters", IRE Trans. on Circuit Theory, Vol. CT-3, No. 3, pp. 348-349, September 1960.

12. Goldstein, A. J. and Kuo, F. F., "Multiparameter Sensitivity", IRE Trans. on Circuit Theory, Vol. CT-8, No. 2, pp. 177-178, June 1961.

13. Blostein, M. L., "Sensitivity Considerations in RLC Networks", Doctoral Dissertation, University of Illinois, August 1963.

14. Lee, S. C., "Network Synthesis to Minimize Multiparameter Sensitivity", Coordinated Science Laboratory Report R-262, University of Illinois, August 1965.

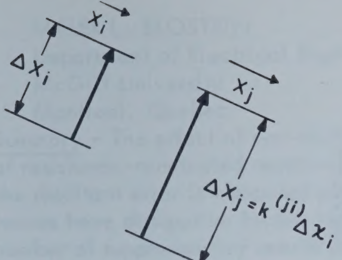

x_i
Δx_i
x_j
Δx_j = k^(ji) Δx_i

FIGURE 1. TWO DEPENDENT PARAMETERS Xi, Xj

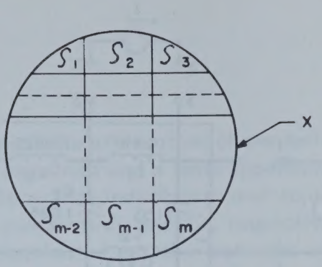

S_1
S_2
S_3
S_m-2
S_m-1
S_m
x

FIGURE 2. SENSITIVITY PARTITION OF X.

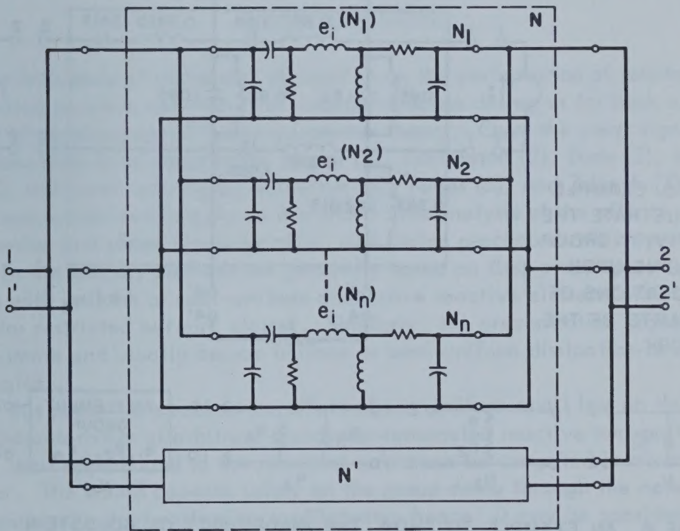

e_i (N_1)
N_1
N
e_i (N_2)
N_2
(N_n)
e_i
N_n
1
1'
2
2'
N'

FIGURE 3. A TYPICAL MULTIPLE TWO-TERMINAL-PAIR NETWORK

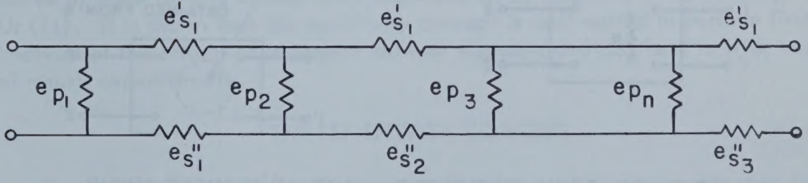

e'_s1
e'_s1
e'_s1
e_p1
e_p2
e_p3
e_pn
e''_s1
e''_s2
e''_s3

FIGURE 4. EXAMPLE OF SENSITIVITY GROUP

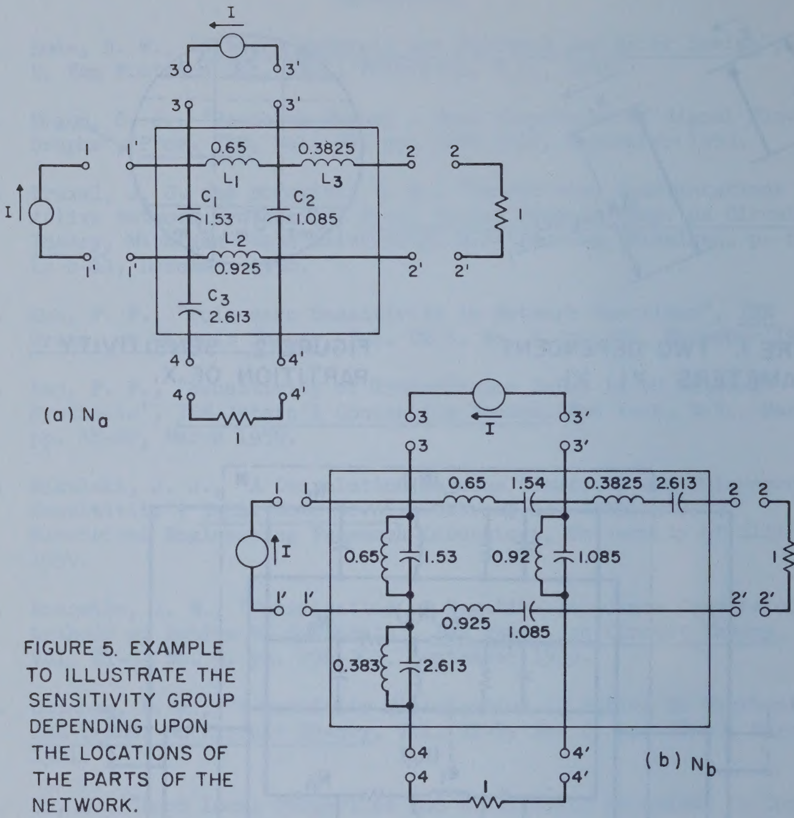

I
3
3'
1
1'
0.65
L1
0.3825
L3
2
2
C1
1.53
C2
1.085
L2
0.925
C3
2.613
I
1'
1'
2'
2'
1
4
4'
(a) N_a
1
3
3'
I
0.65
1.54
0.3825
2.613
0.65
1.53
0.92
1.085
0.925
1.085
0.383
2.613
4
4'
(b) N_b
1

FIGURE 5. EXAMPLE TO ILLUSTRATE THE SENSITIVITY GROUP DEPENDING UPON THE LOCATIONS OF THE PARTS OF THE NETWORK.

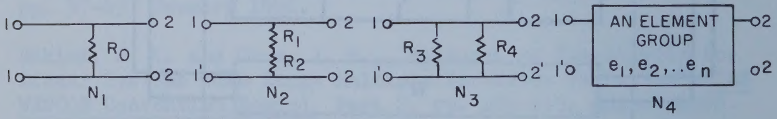

1
2
R_0
N_1
R_1
R_2
N_2
1'
R_3
R_4
N_3
2'
AN ELEMENT GROUP
$e_1, e_2, .. e_n$
N_4

FIGURE 6. AN EXAMPLE TO SHOW THE INADEQUACY OF GOLDSTEIN AND KUO'S DEFINITION.

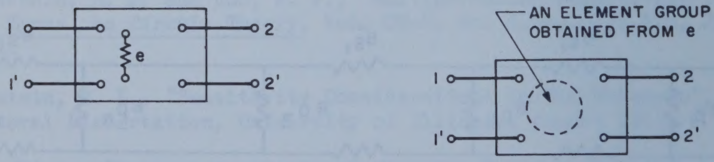

1
1'
e
2
2'
AN ELEMENT GROUP OBTAINED FROM e

FIGURE 7. SPLITTING AN ELEMENT e INTO AN ELEMENT GROUP.

ON THE EFFECT OF LOSS IN FILTER NETWORKS

MAIER L. BLOSTEIN
Department of Electrical Engineering
McGill University
Montreal, Quebec

Summary - The effect of non-uniform small loss on the attenuation characteristics of resistance-terminated reactive two-ports is examined and a least upperbound to the resultant error is computed under the condition that inductances and capacitances have dissipation factors less than fixed constants d_L and d_C, respectively. A number of supplementary results pertaining to terminated reactive networks are obtained including a theorem which states that phase lag through such a network is a monotone increasing function of the magnitude of any reactive element, a least upperbound on the maximum phase error when inductive and capacitive elements have fractional tolerances of $\pm\epsilon_L$ and $\pm\epsilon_C$, respectively, and a first order approximation to the actual attenuation at infinite loss points when loss is placed in series or shunt arms of otherwise arbitrary structures.

INTRODUCTION

The influence of incidental dissipation on the performance of resistance-terminated reactive networks is a classical problem dating as far back as the early period of development of image-parameter theory. Over the years significant contributions have been reported by Mayer (1), Darlington (2), Bode (3), and Guillemin (4), and more recently by Belevitch (5), Temes (6), and Zdunek (7). Studies have been conducted largely on two fronts: the analysis of the effects of loss on attenuation and phase characteristics, and design procedures which precorrect for loss. In the former, methods are generally based on first-order approximation techniques with uniform or semi-uniform dissipative reactive elements which are usually valid for restricted network classes. Similarly, the precorrection procedures are approximate and usually assume uniform or semi-uniform dissipation in restricted topologies.

In this paper we examine the effect of non-uniform small loss on the attenuation characteristics of arbitrary resistance-terminated reactive two-ports and show how a least upperbound to the resultant errors can be computed in a very simple manner. The bound depends solely on the group delay through the network and on the input and output scattering coefficients; hence, it may be considered an invariant once the four-pole parameters of the reactive network are selected. A number of supplementary results pertaining to terminated reactive networks are obtained which are both of academic and practical interest.

The analysis is formulated in terms of the sensitivity function originally introduced by Bode (8) and subsequently expanded on by Angelo (9), Lynch (10), and Ur (11). It is shown that the sensitivity concept is well-suited to perform first-order analyses of the effects of imperfect network elements and can lead to wide variety of results expeditiously.

THE SENSITIVITY CONCEPT

Consider a network in which elements x_1, x_2, ----- x_b, are subject to fractional changes dx_1, dx_2 ---- dx_b; then, a first-order approximation to the corresponding change in a particular network function F can be described by the expression

$$\Delta F/F \approx \sum_{i=1}^{b} S(F\,, x_i)\,\frac{dx_i}{x_i} \qquad (1)$$

where

$$S(F\,, x_i) \triangleq \frac{\partial \ln F}{\partial \ln x_i} = \frac{x_i}{F}\,\frac{\partial F}{\partial x_i}$$

The function $S(F, x_i)$ is defined as the sensitivity of F with respect to x_i . As we shall see, it is a real rational function of the complex frequency variable p and retains singularities wherever F (p) has either a pole or a zero.

Generally speaking x_i represents a real parameter such as R, L, or C, and dx_i/x_i represents a real fractional change in x_i . Under these circumstances the fractional change in the magnitude characteristic is given by (ref. 9),

$$\frac{\Delta\,|F\,(j\omega)|}{|F\,(j\omega)|} \approx \sum_{i=1}^{b} \left[\mathrm{Re}\ \ S\,(F, x_i\,, j\omega)\right] \frac{dx_i}{x_i} \qquad (2)$$

and the net error in the phase characteristic by

$$\Delta \arg F\,(j\omega) \approx \sum_{i=1}^{b} \left[\mathrm{Im}\ \ S\,(F, x_i\,, j\omega)\right] \frac{dx_i}{x_i} \qquad (3)$$

For these reasons Re $[S\,(F, x_i, j\omega)]$ is often referred to as the magnitude sensitivity and Im $[S\,(F, x_i, j\omega)]$ as the phase sensitivity.

The concept of delay sensitivity can also be formulated easily in terms of the sensitivity function. Here, recall that group delay $\tau(\omega)$ can be found from the formula

$$\tau\,(\omega) = -\mathrm{Re}\left[\frac{d \ln F\,(j\omega)}{d\,(j\omega)}\right] \qquad (4)$$

If we define

$$G\,(p) = \frac{d \ln F}{dp} \qquad (5)$$

then the change in G (jω) in response to a fractional change dx_i/x_i becomes

$$\Delta G\,(j\omega) \approx \left[x_i\ \frac{\partial}{\partial x_i}\ \left(\frac{\partial \ln\ F}{\partial\ p}\right)\Bigg|_{p\,=\,j\omega}\right] \frac{dx_i}{x_i} \qquad (6)$$

Since under most circumstances we may write

$$x_i\ \frac{\partial}{\partial x_i}\left(\frac{\partial \ln F}{\partial\ p}\right) = \frac{\partial}{\partial p}\left(x_i\ \frac{\partial \ln F}{\partial\ x_i}\right)$$

it follows that

$$\Delta G\,(j\omega) \approx \left[\frac{\partial}{\partial p}\ \ S\,(F, x_i, p)\Big|_{p\,=\,j\omega}\right] \frac{dx_i}{x_i} \qquad (7)$$

and

$$\Delta \tau\,(\omega) = -\,\frac{\partial}{\partial \omega}\left[\mathrm{Im}\ \ S(F, x_i, j\omega)\right] \frac{dx_i}{x_i} \qquad (8)$$

Thus the derivative of the imaginary part of the sensitivity function with respect to ω can be referred to as the delay sensitivity.

Only slight manipulation is necessary to formulate the effects of incidental dissipation in terms of the sensitivity function. Here it is convenient to regard the nominal value of a reactive element as $p\,\ell_i$ (for inductances) and observe that

$$S(F, p\ell_i) = S(F, \ell_i)$$

Then, the change in the value of an element due to dissipation can be taken as r_i where the dissipation factor of the inductance is given by $d_L = r_i/\ell_i$. As a result, a first approximation to the relative change in the network function can be written as

$$\frac{\Delta F}{F} \approx \frac{r_i}{p\ell_i} \; S(F, \ell_i) \tag{9}$$

Since

$$\frac{\Delta F(j\omega)}{F(j\omega)} \approx d \ln |F(j\omega)| + j \arg F(j\omega) \tag{10}$$

it follows that

$$\frac{\Delta |F(j\omega)|}{|F(j\omega)|} \approx \frac{1}{Q_i}\left[\operatorname{Im} \; S(F, \ell_i, j\omega)\right] \tag{11}$$

where $Q_i = \omega\ell_i/r_i$ and represents the quality factor of the coil. Hence the effect of a real element perturbation on the phase curve is identical to the effect of loss on the magnitude characteristic. In this characterization, the role of fractional tolerance is analagous to inverse quality factor. The influence of incidental dissipation on phase and delay can be treated similarly.

In summary, the sensitivity function defined in (1) is a convenient artifice for studying the effects of both real parameter variations and incidental dissipation on the various types of frequency domain response characteristics normally encountered in practice. Any analytic properties of $S(F, x_i)$ which can be deduced will provide, therefore, insight into a variety of problems relating to influences of imperfect or inexact network elements.

SENSITIVITY FUNCTIONS IN RESISTANCE-TERMINATED LOSSLESS NETWORKS

An important advantage of the sensitivity function is that it can be characterized conveniently in terms of network determinants and cofactors. Now, any network function can be written as $F = W/T$ where W and T represent cofactors or determinants of a system of mesh or nodal equations. Further, W and T are homogeneous polynomials in the network elements and are linear in each x_i. We may write then,

$$F = \frac{W^o + x_i W'}{T^o + x_i T'} \tag{12}$$

where W^o and T^o represent W and T, respectively, with $x_i = 0$, and W' and T' are minors of W and T, respectively. In lieu of (12) it follows readily that (ref.10)

$$S(F, x_i) = \frac{T^o}{T} - \frac{W^o}{W} \tag{13}$$

Equation (13) is a particularly interesting formulation of $S(F, x_i)$ as it places in evidence a number of useful properties. For example, poles and zeros of F (p) are clearly singularities of $S(F, x_i)$ but the converse is not necessarily true since W and T may contain common factors ; the effects of pole displacements are described entirely by T^o/T whereas zero displacements are taken into account by W^o/W.

More importantly T^o/T can be characterized quite simply by noting

$$\frac{T^o}{T} = \frac{1/x_i}{1/x_i + T'/T^o} \tag{14}$$

If $x_i = g_i$, a conductive element, and T refers to a nodal determinant, T'/T^o represents the impedance facing g_i and then T^o/T is a positive real function. Alternatively for capacitive x_i it is convenient to write

$$\frac{T^o}{T} = \frac{1/pc_i}{1/pc_i + T'/T^o} \tag{15}$$

so that T^o/T has the properties of a "negative imaginary-part" function, i.e. the imaginary part of the function is non-positive in the closed right half plane. In a similar fashion, when x_i refers to reciprocal inductance, T^o/T is a positive imaginary-part function. The significance of these properties will be demonstrated in the subsequent discussion.

Generally, W^o/W does not lend itself to an equally simple characterization. However, if the discussion is confined to transfer functions of arbitrary ladder structures (see Figure 1) similar results are valid. This is evident from the following theorem which is a special case of an earlier theorem by Lin and Siskind (12).

Theorem I - In an arbitrary ladder structure where the transfer function of interest is $F = W/T$, W may be factored as $W = W_1 W_2 \text{-----} W_{2n}$, where W_{2k+1} is the nodal determinant of the (2k+1)th series arm subnetwork with external terminals open-circuited ; and W_{2k} is the nodal determinant of the 2k th shunt arm subnetwork with external terminals short-circuited.

Suppose element x_i is contained in the k th subnetwork of the ladder ; then, from Theorem I $W^o/W = W_k^o/W_k$ and since W_k represents a nodal determinant, W_k'/W_k^o represents a driving point impedance, so that the arguments of the preceding paragraph are equally valid for W^o/W.

A particularly interesting property of $S(F, x_i)$ is revealed if, in addition, the ladder is assumed reactive (with resistive terminations). Under these conditions it is evident that W_k'/W_k^o must be a reactance function ; hence (when x_i represents a capacitive element).

$$\begin{aligned} \frac{W^o}{W} &= \frac{1/pc_i}{1/pc_i + W_k'/W_k^o} \\ &= (\text{reactance function}) / p \end{aligned} \tag{16}$$

Consequently W^o/W is purely real on the j-axis and since T^o/T is a negative imaginary-part function, we have

$$\operatorname{Im} S(F, c_i , j\omega) \leq 0 \tag{17}$$

for all ω and every c_i . A similar analysis is valid for inductances ; in this case it is necessary to use the relation

$$S(F, \ell_i) = - S(F, 1/\ell_i)$$

which follows from the logarithmic character of the sensitivity function. In summary, we have the following theorem :

Theorem II - In a resistance-terminated reactive ladder the sensitivity of a transfer function between input and output terminals with respect to an L or C is a negative imaginary-part function.

Physically the above theorem implies that increasing the value of any L or C in the ladder necessarily increases the phase lag through the network at every value of ω. Although the result is derived on the basis of first-order theory, the network remains a reactive ladder even after an element perturbation takes place ; thus, in effect, Theorem II states that phase lag is a monotone increasing function of the magnitude of a reactive element. Alternatively, in the light of Equation (11), we may say that the addition of small loss to any reactive element increases the attenuation at each ω. This result is clearly first-order since on the addition of loss the network ceases to be lossless.

The question immediately arises as to whether the restriction to ladder topologies is necessary for the validity of the theorem. The following argument shows this not to be the case : recall that (4) enables the group delay function to be written as

$$\tau(\omega) = -\mathrm{Re}\left[\frac{dT/dp}{T} - \frac{dW/dp}{W}\right]\Bigg|_{p=j\omega} \tag{18}$$

Since the quadipole is reactive, W is either even or odd so that dW/dp must be either odd or even. The second term on the right hand side of (18) must vanish and we obtain the known result that group delay is independent of W, i.e. of the location of transmission zeros. Equations (8) and (13) then force the conclusion

$$\frac{\partial}{\partial\omega}\left[\mathrm{Im}\,\frac{W^o}{W}\Bigg|_{p=j\omega}\right] = 0$$

which in turn implies that $\mathrm{Im}\left[(W^o/W)\,|\,p = j\omega\right] = 0$ since W and W^o are real polynomials in p. W^o/W is then purely real on the j-axis and Theorem II can be generalized as follows :

Theorem III - In a resistance-terminated lossless network, the phase lag between output and input terminals is a monotone increasing function of the magnitude of any element.

Implications concerning the effects of parasitic-loss and elemental variations on magnitude and phase characteristics, respectively, follow accordingly.

ERROR BOUNDS DUE TO IMPERFECT ELEMENTS

Theorem III can be used effectively to deduce results of practical consequence concerning the effects of loss in filter networks. The point of departure for this discussion is two relationships between sensitivity functions in resistance-terminated reactive networks which have been described in an earlier paper by this author (13). These are summarized below :

$$\sum_{i=1}^{\ell} \mathrm{Im}\, S(F, \ell_i, j\omega) = -\frac{1}{2}\left\{\omega\tau(\omega) + \mathrm{Im}\left[\frac{\rho_1(j\omega) + \rho_2(j\omega)}{2}\right]\right\} \tag{19}$$

$$\sum_{i=1}^{c} \mathrm{Im}\, S(F, c_i, j\omega) = -\frac{1}{2}\left\{\omega\tau(\omega) - \mathrm{Im}\left[\frac{\rho_1(j\omega) + \rho_2(j\omega)}{2}\right]\right\} \tag{20}$$

ℓ and c refer to the total number of inductances and capacitances in the network, respectively, and ρ_1 and ρ_2 represent the input and output scattering coefficients, respectively, normalized to the terminations.

If we define $\Delta\alpha(\omega) = -\Delta|F(j\omega)| \,/\, |F(j\omega)|$ as the error in the attenuation function (the negative sign is present to transform a gain into attenuation) and note the negative imaginary-part character of $S(F, x_i, j\omega)$, it follows from Equation (11)

$$[\Delta\alpha(\omega)]_{max} = -\frac{1}{\overline{Q}_L}\sum_{i=1}^{\ell} \operatorname{Im} S(F, \ell_i, j\omega) - \frac{1}{\overline{Q}_C}\sum_{i=1}^{c} \operatorname{Im} S(F, c_i, j\omega) \qquad (21)$$

where $\overline{Q}_L$ and $\overline{Q}_C$ represent the minimum values of Q-factors of the various coils and capacitors, respectively. Insertion of (19) and (20) in (21) yields

$$[\Delta\alpha(\omega)]_{max} = \frac{1}{2}\left[\bar{d}_L + \bar{d}_C\right]\tau(\omega) + \frac{1}{2}\left[\bar{d}_L - \bar{d}_C\right]\operatorname{Im}\left[\frac{\rho_1(j\omega) + \rho_2(j\omega)}{2\omega}\right] \qquad (22)$$

Therefore, as long as the inductors and capacitors possess dissipation factors less than $\bar{d}_L$ and $\bar{d}_C$, respectively, the maximum deviation of the attenuation characteristic cannot exceed, to a first approximation, the expression in (22) at any value of ω. Also, since the Q's are positive, this maximum deviation is always positive, i.e. represents an increase in attenuation.

It is important to emphasize that the relationships deduced here are first-order approximations and precise limits of applicability are usually difficult to ascertain. Qualitatively, these results are seldom satisfactory in the immediate vicinity of j-axis poles or zeros. This is amply illustrated in Figure 2 where the continuous line represents the attenuation characteristic of a lossless filter near a j-axis transmission zero. The "attenuation pole" is the result of a shunt arm consisting of a series of LC combination (Figure 3). Introduction of loss in this arm produces the dotted curve and clearly between A and B the attenuation of the filter is decreased ! What must be understood here is that for any ω there is always a sufficiently small dissipation factor for which these results are useful, but as ω approaches a transmission zero the permitted range of loss factors becomes infinitesimal.

The treatment in the vicinity of transmission zeros can be handled more effectively in reactive ladders by calculating the shift in the zero location. This may be computed by recalling that if a partial fraction expansion of Equation (9) is carried out, the residue at any simple singularity of $S(F, x_i)$ due to W represents, to a first approximation, the negative of the displacement of the corresponding zero of F (p) (11). From the derivation of Theorem II it is apparent that $(r_i/p\ell_i)\, W^o/W$ is a reactance function and thereby possesses only positive real residues. We may then formulate,

Theorem IV - The introduction of loss in any arm of a resistance-terminated reactive ladder necessarily shifts the transmission zero produced by this arm into the left half plane, parallel to the j-axis, to an extent proportional to the amount of loss.

The extent of the shift is readily calculable from the product of r_i and the residue of an easily obtainable one-port LC impedance function. As an example, the transmission zero produced by a shunt arm such as in Figure 3 is perturbed by the extent

$$\Delta p_i \approx \frac{r_i}{2\ell_i} \qquad (23)$$

due to inductive loss (additional capacitive loss would simply provide an added term). The perturbed attenuation characteristic can then be computed in the vicinity of the transmission zero by assuming the new location for the zero but neglecting corresponding pole perturbations. In particular, if the zero is located originally at $p = j\omega_o$, the attenuation at the infinite loss point is

$$\alpha(\omega_o) = \log \left| \left\{ \frac{d_i}{2} \left. \frac{F(p)}{(p - j\omega_o)} \right|_{p = j\omega_o} \right\} \right| \tag{24}$$

These results can be extended to series or shunt arms of any complexity with arbitrary (but small) loss in any element.

DISCUSSION OF RESULTS

Of the various results presented, perhaps the most significant is Equation (22) where a first-order approximation is given for the maximum error introduced into an attenuation characteristic due to non-uniformly distributed loss. This result generalizes earlier work by Belevitch (5) in two ways. First, he assumed a symmetrical two-port and secondly his approximation was based upon a semi-uniform distribution of losses. The results reported here remove the symmetry restriction (they are equivalent to those of Belevitch for the symmetric semi-uniformly dissipative case) and show that Equation (22) represents a least upperbound for the non-uniform case provided maximum values of dissipation constants are used in the formula. Similar remarks follow with regard to computing attenuation at infinite loss points, where previous work has concentrated principally on image-parameter filters (4,5). It is useful to observe that Equation (22) could be anticipated, at least in the case of ladder structures, from Mayer's theorem on the effects of uniform dissipation (see Bode pp.220) and Equations (17) and (56) in Temes' work. Theorem IV was originally deduced by Temes in a somewhat different manner but the proof offered here leads more expeditiously to the determination of the magnitude of the zero shift.

From the design viewpoint, Equation (22) can be used in situations where losses are not excessive and it is desirable to avoid predistortion techniques. In fact, $\bar{d}_C$ and $\bar{d}_L$ can be computed in order to constrain the error between specified limits. This may be particularly valuable wherever loss components are subject to climatic variations. Similar remarks follow in relation to computing the minimum permitted loss at attenuation poles.

Useful results can also be obtained concerning the effect of element tolerances on phase error. Thus, let n_i represent the nominal value of the i th inductance ℓ_i. If we assume

$$n_i - \epsilon_L \leq \ell_i \leq n_i + \epsilon_L \qquad i = 1, 2, --- \ell$$

where $\epsilon_L > 0$, then the maximum error in the phase curve can be written as

$$\Delta \phi(\omega) = \epsilon_L \sum_{i=1}^{\ell} \operatorname{Im} S(F, \ell_i, j\omega)$$

$$= -\frac{\epsilon_L}{2} \left\{ \omega \tau(\omega) + \operatorname{Im} \left[\frac{\rho_1(j\omega) + \rho_2(j\omega)}{2} \right] \right\} \tag{23}$$

The effects of capacitive variations are treated similarly and provide an additional term in (23). The practical applications of (23) are apparent.

The problem of bounding the phase error due to incidental dissipation, which is analagous to bounding the attenuation error due to real parameter variations, remains open and subject to further investigation. Also, the problem of bounding delay errors for either type of imperfection remains unsolved.

REFERENCES

1. H.F. Mayer, "Veber die Daempfung von Siebketten im Durchlaessigkeitsbereich", ENT Vol.2, pp. 335-338 ; October, 1925.
2. S. Darlington, "Synthesis of reactance 4-poles which produce prescribed insertion loss characteristics", J. of Math. and Phys., Vol.18, pp. 257-353 ; September, 1939.
3. H. Bode, "Network Analysis and Feedback Amplifier Design", D. Van Nostrand Co. Inc., New York, N.Y., pp. 216-223 ; 1945.
4. E.A. Guillemin, "Effect of Dissipation in Filters", Electronics, Vol. 19, pp. 130-5 ; October, 1946.
5. V. Belevitch, "Theorie des Circuits de Telecommunication", Gauthier-Villars, Paris, Frànce, pp. 205-212 ; 1957.
6. G. Temes, "First-Order Estimation and Precorrection of Parasitic Loss Effects in Ladder Filters", IRE Trans. on Circuit Theory, Vol. CT-9, pp. 385-400 ; December, 1962.
7. J. Zdunek, "General Design of Ladder Networks with arbitrarily distributed uniform losses between coils and capacitors", Proc. of IEE, Vol. 110, pp. 295-302 ; February, 1963.
8. H. Bode, op. cit., pp. 52.
9. E.J. Angelo, "Design of Feedback Systems", Research Report R-449-55 PIB-379, Polytechnic Institute of Brooklyn ; 1955.
10. W.A. Lynch, "A Formulation of the Sensitivity Function", IRE Trans. on Circuit Theory, Vol. CT-4, p. 289 ; September, 1957.
11. H. Ur, Discussion on paper by S.Y. Huang entitled "Sensitivity of the Poles of Linear Closed Loop Systems", AIEE Transactions on Applications and Industry, Vol. 38, pp. 182-187 ; September, 1958.
12. P.M. Lin and R.P. Siskind, "A Simplified Cascade Synthesis of RC Transfer Functions", IEEE Trans. on Circuit Theory, Vol. CT-12, pp. 98-107 ; March, 1965.
13. M.L. Blostein, "Bounds on Sensitivity in RLC Networks", Proc. of Allerton Conf. on Circuit and System Theory, Vol. I, pp. 488-502 ; 1963.
14. J.H. Mole, "Filter Design Data for Communication Engineers", E. & F. Spon Ltd., London, pp. 140-160 ; 1952.

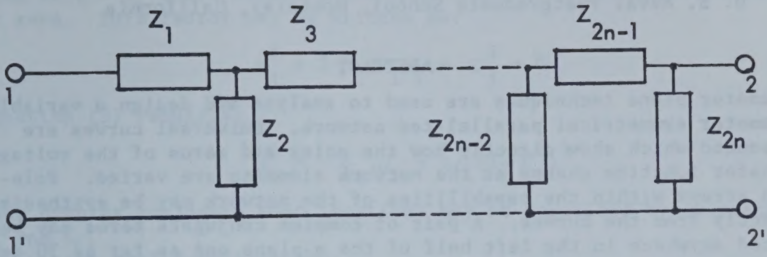

Z_1
Z_3
Z_{2n-1}
1
2
Z_2
Z_{2n-2}
Z_{2n}
1'
2'

FIGURE 1.

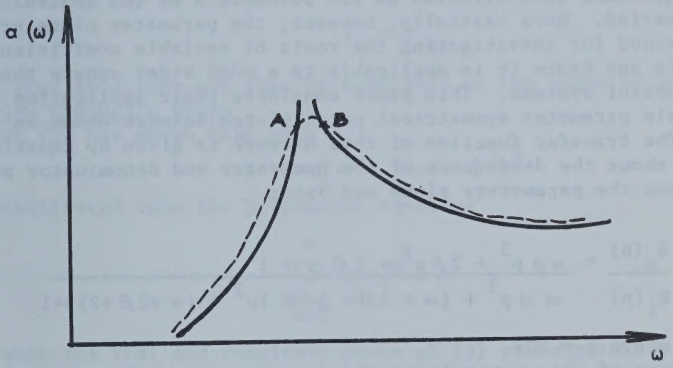


FIGURE 2.

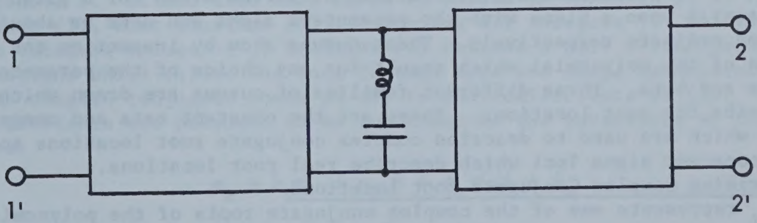


FIGURE 3.

SYMMETRICAL PARALLEL-TEE NETWORK--PARAMETER PLANE ANALYSIS AND SYNTHESIS

F. H. HOLLISTER AND G. J. THALER
U. S. Naval Postgraduate School, Monterey, California

ABSTRACT

Parameter plane techniques are used to analyze and design a variable parameter symmetrical parallel-tee network. Universal curves are presented which show directly how the poles and zeros of the voltage transfer function change as the network elements are varied. Pole-zero arrays within the capabilities of the network may be synthesized directly from the curves. A pair of complex conjugate zeros may be placed anywhere in the left half of the s-plane and as far as 30 degrees from the imaginary axis into the right half of the s-plane by choosing network parameter values from the curves.

INTRODUCTION

Parameter plane techniques evolved from a study of the characteristic equations of variable parameter control systems and of how the roots of such equations were affected as two parameters of the control system were varied. More basically, however, the parameter plane provides a method for investigating the roots of variable coefficient polynomials and hence it is applicable to a much wider sphere than that of control systems. This paper considers their application to the variable parameter symmetrical parallel-tee network shown in Fig. 1. The transfer function of this network is given by Equation (1) which shows the dependence of the numerator and denominator polynomials upon the parameters alpha and beta:

$$\frac{E_o(p)}{E_i(p)} = \frac{\alpha\beta p^3 + 2\beta p^2 + 2\beta p + 1}{\alpha\beta p^3 + (\alpha + 2\beta + 2\alpha\beta)p^2 + (\alpha + 2\beta + 2)p + 1} \tag{1}$$

where $p = RCs$

The objective of this paper is to show how parameter plane techniques can be employed to determine the poles and zeros of this transfer function for any choice of the network parameters or, conversely, how to select parameter values which will result in desired pole or zero locations.

PARAMETER PLANE THEORY

A parameter plane consists of families of curves drawn for a given polynomial upon a plane with the parameters alpha and beta as abscissa and ordinate respectively. These curves show by inspection the roots of the polynomial which result for any choice of the parameters alpha and beta. Three different families of curves are drawn which describe the root locations. These are the constant zeta and omega loci which are used to describe complex conjugate root locations and the constant sigma loci which describe real root locations.

Describing Complex Conjugate Root Locations

If s_i represents one of the complex conjugate roots of the polynomial:

$$\sum_{k=0}^{n} a_k s^k \tag{2}$$

then at $s = s_i$ there exists a quadratic factor of (2) which is equal to zero. This factor may be written as:

$$s_i^2 + 2\zeta_i \omega_i s_i + \omega_i^2 = 0 \tag{3}$$

Equation (3) requires:

$$s_i = -\zeta_i \omega_i + \omega_i (1-\zeta_i^2)^{\frac{1}{2}} \tag{4}$$

Thus complex conjugate root locations may be described by quantities ζ and ω.

Describing Real Root Locations

If s_j represents a simple real root of the polynomial (2), then at $s=s_j$ there is a first degree factor of (2) which is equal to zero. This may be written as:

$$s_j - \sigma_j = 0 \tag{5}$$

Equation (5) may be written:

$$s_j = \sigma_j \tag{6}$$

Real root locations may thus be described by the quantity σ.

Complex Conjugate Roots

Siljak (1) has shown that when

$$s = -\zeta\omega + j\omega(1-\zeta^2)^{\frac{1}{2}} \tag{7}$$

is substituted into the polynomial equation

$$\sum_{k=0}^{n} a_k s^k = 0 \tag{8}$$

and when the real and imaginary parts of (8) are separately equated to zero, the following relationship holds:

$$\sum_{k=0}^{n} (-1)^k a_k \omega^k U_{k-j}(\zeta) = 0 \quad \text{for } j = 0,\ 1 \tag{9}$$

The functions $U_i(\zeta)$ are Chebyshev functions which are tabulated in Appendix I. It has also been shown by Hudson in Appendix II of reference (2), that Equation (9) is true for all integers j. For any two integral values of j, Equation (9) defines a pair of independent equations which relate the root factors zeta and omega associated with complex conjugate roots of (8) to the coefficients of (8). The coefficients are presumed to be functions of the parameters of the form:

$$a_k = b_k\alpha + c_k\beta + d_k\alpha\beta + f_k \tag{10}$$

where: b_k, c_k, d_k, and f_k are real constants
α and β are real and adjustable

When coefficients of the form (10) are substituted into Equation (9) and any two integral values of j are chosen, two independent equations relating the root factors zeta and omega to the parameters alpha and beta result:

$$F(\alpha, \beta, \zeta, \omega) = 0 \tag{11}$$

$$G(\alpha, \beta, \zeta, \omega) = 0 \tag{12}$$

When omega is eliminated between (11) and (12), the equation of the constant zeta loci is obtained:

$$Z(\alpha, \beta, \zeta) = 0 \tag{13}$$

When zeta is eliminated between (11) and (12) the equation of the constant omega loci is obtained:

$$W(\alpha, \beta, \omega) = 0 \tag{14}$$

For any fixed value of $\zeta = \zeta_i$, Equation (13) defines a plane algebraic curve in the parameter plane which specifies the relation between the parameters alpha and beta which is necessary if Equation (8) is to have a pair of complex conjugate roots whose root factor ζ is equal to ζ_i. Likewise for any fixed value of $\omega = \omega_i$, Equation (14) specifies the relation between the parameters which is necessary if Equation (8) is to have a pair of complex conjugate roots whose root factor ω is equal to ω_i. Assume that the $\zeta = \zeta_i$ curve intersects the $\omega = \omega_i$ curve at the coordinates $\alpha = \alpha_i$ and $\beta = \beta_i$ in the parameter plane. When the parameters alpha and beta are adjusted so that $\alpha = \alpha_i$ and $\beta = \beta_i$, the polynomial (8) will have the factor:

$$s^2 + 2\zeta_i \omega_i s + \omega_i^2$$

In this manner parameter values may be chosen from the curves which will result in desired complex conjugate roots.

Real Roots

When $s = \sigma$ and the coefficient form (10) are substituted into Equation (8), the equation of the constant sigma loci results:

$$S(\alpha, \beta, \sigma) = [\sum_{k=0}^{n} b_k \sigma^k]\alpha + [\sum_{k=0}^{n} c_k \sigma^k]\beta +$$

$$[\sum_{k=0}^{n} d_k \sigma^k]\alpha\beta + \sum_{k=0}^{n} f_k \sigma^k = 0 \tag{15}$$

For any fixed value of σ, say $\sigma = \sigma_j$, Equation (15) defines an algebraic curve in the parameter plane which specifies the relation between alpha and beta which is necessary if Equation (8) is to have a real root at $s = \sigma_j$. For any choice of alpha and beta, say α_j and β_j, the coefficients, and hence the roots, of (8) are determined. Thus, if this choice of parameters results in Equation (8) having r distinct real roots, then r constant sigma curves will pass through the point (α_j, β_j) in the parameter plane. It is possible to draw constant sigma curves for selected values of sigma which will define all of the real roots of Equation (8) for any choice of the parameters alpha and beta.

APPLICATION TO THE SYMMETRICAL PARALLEL-TEE NETWORK

Separate frequency normalized parameter planes will be constructed for the poles and zero of Equation (1). These parameter planes will show by inspection the parameter values alpha and beta necessary to place selected singularities of Equation (1) at desired locations in the p-plane. The range of alpha and beta values is arbitrarily chosen as:

$$0 < \alpha, \beta < 2.5 \qquad (16)$$

Poles of the Transfer Function

Since the symmetrical parallel-tee network is passive RC, all of the transfer function poles lie on the negative real axis in the p-plane. Let σ denote the value of p at which such a pole occurs. For the denominator polynomial of Equation (1), Equation (15) becomes:

$$S(\alpha, \beta, \sigma) = (\sigma^3 + 2\sigma^2)\alpha\beta + (\sigma^2 - \sigma)\alpha + 2(\sigma^2 + \sigma)\beta + (2\sigma + 1) = 0 \qquad (17)$$

Equation (17) represents a family of hyperbolas in the parameter plane. Each hyperbola corresponds to a different value of sigma. Since there are three real poles of (1), it follows that three hyperbolas will pass through any selected point in the parameter plane, thus specifying the three poles which occur for this selection of alpha and beta. Figure 2 shows Equation (17) for selected values of sigma.

Real Zeros of the Transfer Function

Since the parameters alpha and beta are presumed real, it follows that Equation (1) has at least one real zero. As before, let $p = \sigma$ denote such a real zero. Equation (15) becomes:

$$S(\alpha, \beta, \sigma) = \sigma^3\alpha\beta + 2(\sigma^2 + \sigma)\beta + 1 = 0 \qquad (18)$$

Equation (18) also represents a family of hyperbolas which is displayed in Figure 3 for selected values of sigma.

Complex Conjugate Zeros of the Transfer Functions

For the numerator polynomial of Equation (1) and with j = 0 and j = 3, Equation (9) produces the independent pair of equations:

$$-\alpha\beta\omega^3 U_3(\zeta) + 2\beta\omega^2 U_2(\zeta) - 2\beta U_1(\zeta) + U_0(\zeta) = 0 \qquad (19)$$

$$-\alpha\beta\omega^3 U_0(\zeta) + 2\beta\omega^2 U_{-1}(\zeta) - 2\beta U_{-2}(\zeta) + U_{-3}(\zeta) = 0 \qquad (20)$$

When the appropriate functions of zeta are substituted into (19) and (20) for the Chebyshev functions, and when omega is eliminated between (19) and (20) by Sylvester's Dyalitic Method of Elimination(3), the equation of the constant zeta loci is obtained:

$$Z(\alpha, \beta, \zeta) = -\alpha\beta^3(4\zeta^2 - 1)^3[(-32\alpha\zeta^2 + 16)\beta^2 + (64\alpha\zeta^4 + 16\alpha\zeta^2 - 32\zeta^2 - 8\alpha)\beta + (-64\alpha^2\zeta^6 + 48\alpha^2\zeta^4 - 12\alpha^2\zeta^2 + \alpha^2)] = 0 \qquad (21)$$

Constant zeta loci are displayed in Figure 3 for selected values of zeta. Observe that negative values of zeta correspond to right half plane complex conjugate zeros.

By letting j = 0, 1, and 2 and making use of the relation $U_{-k} = -U_k$, Equation (9) generates a set of three linear and homogeneous equations in the quantities U_1, U_2, and U_3. In order for a non-trival solution to exist, it is necessary that the determinant of the coefficient matrix of this set of equations vanish. This condition produces the equation of the constant omega loci:

$$W(\alpha,\beta,\omega) = \alpha\beta\omega^3(\alpha^2\beta^2\omega^6-2\alpha\beta^2\omega^4-2\beta\omega^2-1) = 0 \quad (22)$$

Constant omega loci are displayed in Figure 3 for selected values of omega.

THE NATURE OF THE ZEROS

Inspection of the parameter plane for the transfer function zeros reveals the following information:

The positive alpha axis coincides with the $\zeta = -\frac{1}{2}$ locus indicating that zeros may be placed as far as 30 degrees from the imaginary axis into the right half of the p-plane.

The $\zeta = 0$ locus is a straight line whose equation is $\alpha = 4\beta$. This indicates that if one is interested only in those cases where a pair of conjugate zeros is placed on the imaginary axis, then this linear relation may be substituted into the network equations, thus reducing the problem to a single dimensional one.

The $\zeta = 1$ locus separates the region of three real zeros from the region of two complex conjugate zeros and one real zero. This locus is also the locus of repeated real zeros.

The $\zeta = \frac{1}{2}$ locus coincides with the $\omega = 1$ locus indicating that an additional degree of freedom exists which may be used to locate additional poles or zeros in addition to locating a pair of complex conjugate zeros at $\zeta = \frac{1}{2}$, $\omega = 1$.

CONCLUSION

The authors have shown how a network with two adjustable parameters may be investigated by parameter plane methods. From the curves obtained, parameter values can be selected which will result in desired pole or zero locations or, conversely, poles and zeros may be determined which result for selected parameter values. The parameter plane technique is quite general and provides information about the roots of polynomials whose coefficients are functions of two variable parameters, which is equivalent to information provided by root locus methods for polynomials involving a single variable parameter.

ACKNOWLEDGEMENT

The authors wish to acknowledge the work done by Barker and Rosenstein (4) in analyzing the symmetrical parallel-tee network with variable parameters. Their work, while complete and well done, was somewhat complicated by the inherent single parameter capabilities of the root locus technique which they employed.

APPENDIX I

Table of the Chebyshev Functions $U_i(\zeta)$:

i	$U_i(\zeta)$
0	0
1	1
2	2ζ
3	$4\zeta^2-1$

Relations for the Chebyshev Functions $U_i(\zeta)$:

$$U_{-i}(\zeta) = -U_i(\zeta)$$

$$U_{k+1}(\zeta) = 2\zeta U_k(\zeta) - U_{k-1}(\zeta)$$

REFERENCES

1. Šiljak, D. D., Analysis and Synthesis of Feedback Control Systems in the Parameter Plane, Part I, Linear Continuous Systems, IEEE Transactions on Applications and Industry, Nov. 1964, pp. 449-458.

2. Hollister, F. H., Network Analysis and Design by Parameter Plane Techniques, Dissertation, U. S. Naval Postgraduate School, Monterey, California, 1965.

3. Dickson, L. E., First Course in the Theory of Equations, Wiley and Sons, New York, 1922, pp. 145-147.

4. Barker, A. C., and Rosenstein, A. B., s-Plane Synthesis of the Symmetrical Twin-Tee Network, Proc. of the National Electronics Conference, 1964, Vol. 20, pp. 542-547.

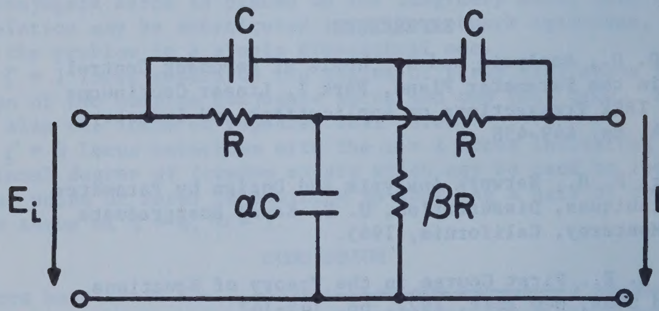
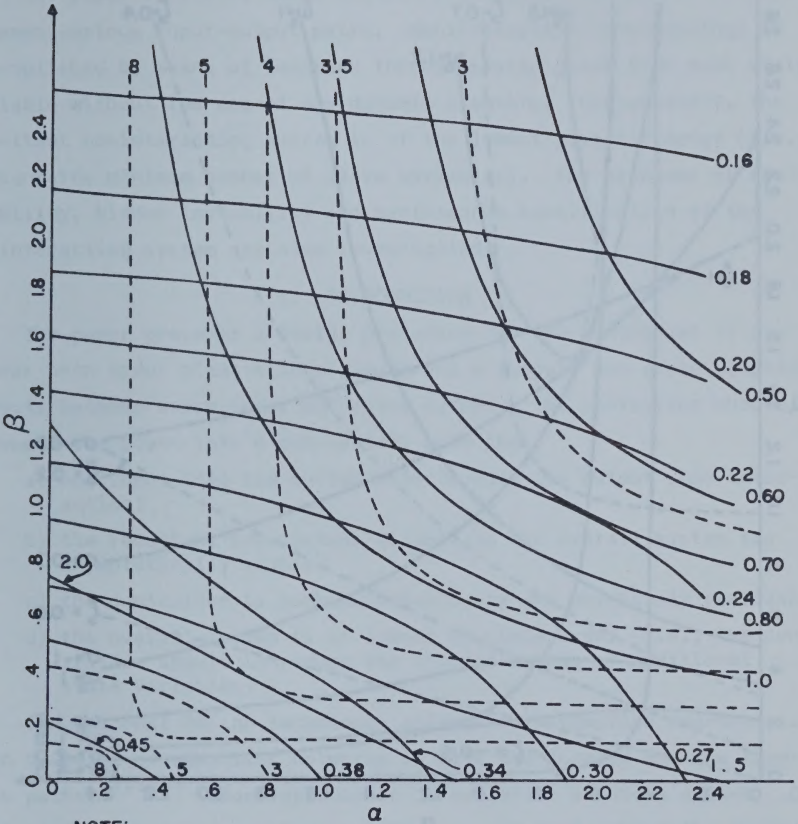

C
C
R
R
Ei
αC
βR
Eo

FIG.1 SYMMETRICAL PARALLEL-TEE NETWORK



NOTE:
CURVE LABELS REFER TO
NEGATIVE REAL ROOT VALUES

FIG.2 POLES OF THE SYMMETRICAL PARALLEL-TEE NETWORK

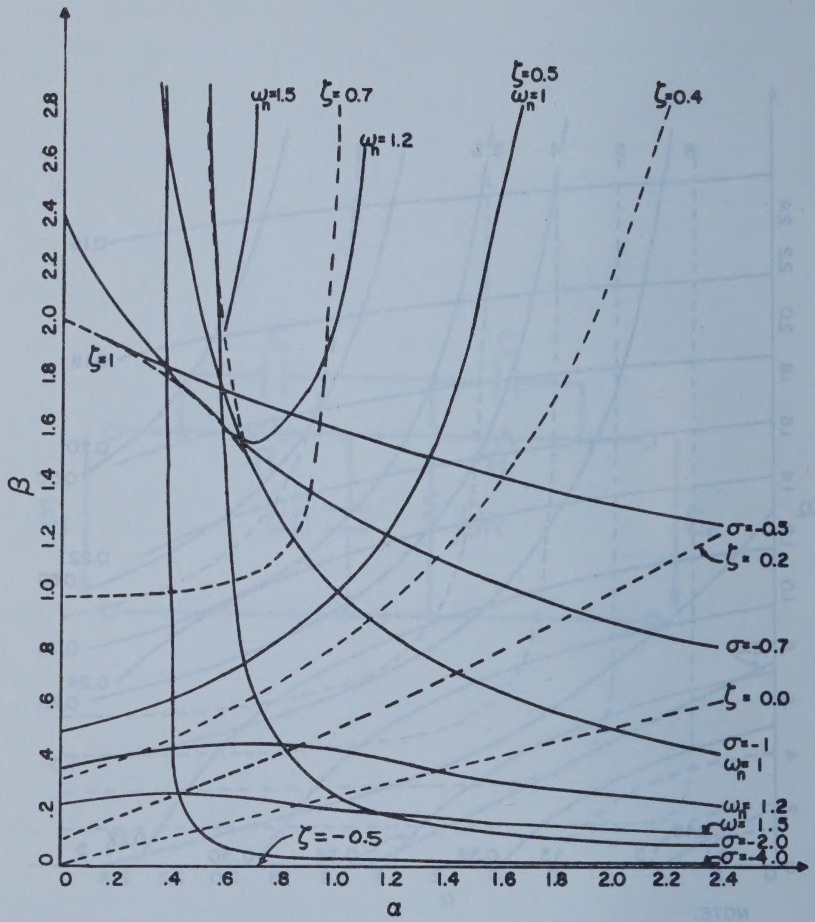


FIG.3 ZEROS OF THE SYMMETRICAL PARALLEL-TEE NETWORK

DECOUPLING OF MULTIVARIABLE SYSTEMS BY MEANS OF STATE VARIABLE FEEDBACK

Z. V. REKASIUS
Department of Electrical Engineering
Northwestern University

ABSTRACT

The paper considers multivariable linear plants with cross-coupling between various input-output pairs. Noninteraction (=decoupling) is accomplished by means of feedback through static gains from each state variable without the use of any dynamic elements. Consequently, the resultant noninteracting system is of the lowest possible order (i.e., it contains minimum number of state variables). The problems of realizability, hidden instability and performance specification of the noninteracting system are also investigated.

I. INTRODUCTION

The paper presents a design procedure for the controller of a linear n-th order plant with m inputs and m outputs and cross coupling effects between every input and every output. The controller should decouple the plant into m sub-systems such that:

a) one input into the system effects only one output (non-interaction)

b) the resultant sub-systems as well as the overall system are asymptotically stable

c) the controller is both mathematically and physically realizable

d) the overall system is of lowest possible order (i.e., the controller should introduce the minimum number of additional state variables)

Conventional design techniques attempt to accomplish non-interaction by means of equalizer networks in both the forward and the feedback paths.(1,2,3) These approaches do not offer a unique answer. One is forced to make more or less arbitrary assumptions about the desired transfer functions of the final noninteracting sub-systems. On the surface these approaches (which represent a direct extension of the closed-loop synthesis of single input-single output systems(4)) appear to offer almost unlimited possibilities in obtaining any desired transfer characteristics. Actually, however, one runs into problems of mathematical and physical realizability. Mathematical unrealizability means that the equations for equalizer networks fail to have a solution. We will show that a necessary and sufficient condition for mathematical realizability is that the plant transfer matrix be non-singular. Physical unrealizability means that the equalizer networks in the controller call for more differentiations than integrations (i.e., the number of zeros in their transfer functions exceeds the number of poles). This happens because one specifies the transfer functions of the overall non-interacting system with insufficient number of poles (provided that the non-interacting system is mathematically realizable).

Even if a non-interacting system is both mathematically and physically realizable, it may turn out to be unstable even if the diagonal elements of the non-interacting transfer matrix have poles only in the left-half of the s-plane[5,6].

Also if one obtains a realizable non-interacting controller it will inevitably call for rather complex equalizer networks. This is because closed-loop synthesis procedures tend to cancel the plant poles by the equalizer zeros and replace them by suitably located equalizer poles.

An attempt to design non-interacting systems with low order equalizers by means of feeding back all the state variables of the plant has been reported by Morgan[5]. Unfortunately, his proposed state variable feedback design yields a non-interacting system only if such system consists of m first order noninteracting sub-systems.

II. PROBLEM STATEMENT

The plant under consideration is represented by the vector differential equation

$$\dot{x} = A\,x + B\,u \tag{2.1}$$

$$y = C\,x \tag{2.2}$$

where n-vector x represents the state of the plant, m-vector u - the input into the plant and m-vector y - the output of the plant. Assuming zero initial state and Laplace transforming the above equations one obtains

$$X(s) = (sI-A)^{-1}\ B\ U(s) \tag{2.3}$$

and

$$Y(s) = C(sI-A)^{-1}\ B\ U(s) \tag{2.4}$$

Defining the plant transfer matrix H(s) by

$$Y(s) \underline{\underline{\Delta}} H(s)\ U(s) \tag{2.5}$$

one gets

$$H(s) = C(sI-A)^{-1} B \tag{2.6}$$

The most general linear controller, containing an equalizer network P(s) in the forward path, an equalizer network Q(s) in the feedback path and incorporating all the state variables into the feedback control law is defined by

$$U(s) = P(s)\,V(s) + Q(s)\,X(s) \tag{2.7}$$

Here the m-vector V(s) represents the Laplace transform of input into the overall system (the m-vector v(t)). Note that if the matrices P and Q in (2.6) are constant this formulation implies linear control with constant feedback of all state variables and no dynamic equalizer networks. Otherwise the equalizer networks P(s) and Q(s) introduce additional state variables. Let these additional state variables be designated by an i-vector $\theta(t)$ and a j-vector $\phi(t)$. Then

$$\dot{\theta} = A_p\,\theta + B_p\,v \tag{2.8}$$

$$\gamma = C_p\theta + D_p\,v \tag{2.9}$$

$$\dot{\phi} = A_q\,\phi + B_q\,x \tag{2.10}$$

$$\delta = C_q\,\phi + D_q\,x \tag{2.11}$$

and

$$u = \gamma + \delta \tag{2.12}$$

$$P(s) = C_p(sI-A_p)^{-1}\,B_p + D_p \tag{2.13}$$

$$Q(s) = C_q(sI-A_q)^{-1}\,B_q + D_q \tag{2.14}$$

(see Fig. 1) Eliminating the variables X(s), U(s), from equations (2.3), (2.4), (2.7) and substituting (2.6) one finds

$$Y(s) = H(s)\left[I-Q(s)\,(sI-A)^{-1}\,B\right]^{-1} P(s)V(s) \tag{2.15}$$

Define the overall system transfer matrix H*(s) by

$$Y(s) = H^*(s)\,V(s)$$

Then

$$H^*(s) = H(s)\,[I-Q(s)\,(sI-A)^{-1}\,B]^{-1}\,P(s) \tag{2.16}$$

or

$$P(s) = [I - Q(s)\,(sI-A)^{-1}\,B]^{-1}\;\;H(s)H^*(s) \tag{2.17}$$

Equations (2.16) and (2.17) indicate that:

> the necessary and sufficient condition for the existence of a non-interacting control for the plant (2.1), (2.2) is that its transfer matrix H(s) be nonsingular.

The necessity of the above result follows immediately from (2.16). For non-interaction H*(s) must be diagonal. Furthermore, H*(s) must be non-singular. Otherwise one of the elements of H*(s), say $h^*_{ii}(s) = 0$ which means that one has no control over the corresponding output $y_i(t)$. Hence, from (2.16), H(s) must be nonsingular. The sufficiency can most easily be shown by choosing

$$Q(s) = 0 \tag{2.18}$$

Then

$$P(s) = H^{-1}(s)\;H^*(s) \tag{2.19}$$

i.e., the non-interacting controller P(s) exists as long as H(s) is nonsingular.

Note that the equalizer specified by (2.19) yields an "open-loop" non-interacting system (certainly not a very desirable solution). One could, of course, design such a non-interacting open-loop system and then close the loops between individual input-output pairs according to conventional design rules for single input-single output systems, without destroying non-interaction. However, (2.19) is more of theoretical interest, since it indicates that for nonsingular plant matrices H(s) one can always achieve physically realizable non-interaction by choosing the diagonal matrix H*(s) with sufficient number of poles in each of its elements. Unfortunately, the transfer function approach does not guarantee that the non-interacting system will not exhibit hidden instability.

In the following sections we will utilize constant gain equalizers and feedback from all state variables. Hence, (2.7) will become

$$u = p\ v + Q\ x \tag{2.20}$$

and the state of the system will be determined by

$$\dot{x} = (A + B\ Q)\ x + B\ P\ v. \tag{2.21}$$

III. NON-INTERACTION BY STATE VARIABLE FEEDBACK

In this section it is assumed that all the plant state variables x_i are available for measurement and that noninteraction is to be achieved without increase in the order of the system. That is, noninteraction is to be achieved by means of constant gain matrices P and Q in (2.7). Let the noninteracting system be

$$\alpha_o\ \dot{y} + \alpha_1\ y = \beta_1\ v \tag{3.1}$$

where α_o, α_1 and β_1 are diagonal m x m matrices. Differentiating (2.2) and substituting (2.21) and (3.1) one obtains

$$\alpha_o\ \dot{y} + \alpha_1\ y = [\alpha_o\ C(A + BQ) + \alpha_1 C]\ x + \alpha_o\ CBPv \tag{3.2}$$

Hence, comparing (3.2) and (3.1) one obtains the following sufficient condition for noninteraction

$$\alpha_o\ C(A + BQ) + \alpha_1\ C = 0 \tag{3.3}$$

and

$$\beta_1 = \alpha_o\ C\ B\ P \tag{3.4}$$

or

$$P = (\alpha_o\ C\ B)^{-1}\ \beta_1 \tag{3.5}$$

and

$$Q = -(\alpha_o \; C \; B)^{-1} (\alpha_1 \; C + \alpha_o \; C \; A) \tag{3.6}$$

The result was first reported by Morgan[5]. Obviously, the approach fails if the matrix C B is singular. If, however, the matrix C B is nonsingular, then the plant can be decoupled into m non-interacting first order sub-systems specified by (3.1) The majority of the realistic problems is such that the number of state variables is much greater than the number of inputs. Hence it is very likely that such plants cannot be decoupled into first-order sub-systems while non-interaction may still be accomplished by decoupling such plants into m sub-systems of higher order. Consequently, the matrix C B will be singular in the majority of realistic problems. A straightforward extension of the approach to nontrivial cases is to specify the noninteracting system as

$$\alpha_o \; y^{(k)} + \alpha_1 \; y^{(k-1)} + \ldots + \alpha_{k-1} \; y + \alpha_k \; y = \beta_1 \; v^{(k-1)} + \beta_2 \; v^{(k-2)} + \beta_{k-1} \; \dot{v} + \beta_k \; v \tag{3.7}$$

where all β_i's and α_i's are diagonal m x m matrices.

The overall transfer matrix then becomes

$$H^*(s) = [\sum_{i=0}^{k} s^{k-i} \; \alpha_i]^{-1} [\sum_{i=1}^{k} s^{k-i} \; \beta_i] \tag{3.8}$$

Repeating the procedure used to arrive at (3.3) and (3.4) one obtains

$$\beta_i = \sum_{j=1}^{i} \alpha_{j-1} \; C(A + BQ)^{i-j} \; B \; P; \qquad i = 1,2,\ldots k \tag{3.9}$$

and

$$\sum_{i=0}^{k} \alpha_i \; C(A + BQ)^{k-i} = 0 \tag{3.10}$$

If the matrix C B is nonsingular, one can choose k=1 and solve (3.8) and (3.9) for P and Q. This yields the previously obtained design equations (3.5) and (3.6). If, however, the matrix C B is singular, one cannot solve (3.9) and (3.10) for P and Q with k=1. Hence one must try k = 2,3,...,etc. to decouple the plant into sub-systems of higher order.

Consider the case where

$$C \; A^i B = 0 \tag{3.11}$$

for all i = 0,1,2,...,k-2. Then (3.10) reduces to

$$\alpha_k C + \alpha_{k-1} \; C \; A + \ldots + \alpha_1 \; C \; A^{k-1} + \alpha_o C \; A^k = - \alpha_o C \; A^{k-1} \; BQ \tag{3.12}$$

and (3.9) becomes

$$\beta_k = \alpha_o \; C \; A^{k-1} \; B \; P; \; \beta_i = 0, \; i = 0,1,2,. \; k-1 \tag{3.13}$$

Hence, if $C A^{k-1}$ is nonsingular, one obtains the noninteracting controller matrices

$$P = (\alpha_o C A^{k-1} B)^{-1} \beta_k \tag{3.14}$$

and

$$Q = -(\alpha_o C A^{k-1} B)^{-1} \sum_{i=0}^{k} \alpha_i C A^{k-i} \tag{3.15}$$

Example 1:

Consider the plant shown in Fig. 2a. Here

$$A = \begin{matrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -2 \end{matrix} \qquad B = \begin{matrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{matrix} \qquad C = \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{matrix} \qquad CB = \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}$$

and CAB is non-singular. Hence k = 2.

Let

$$\alpha_o = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \qquad \alpha_1 = \begin{matrix} 2 & 0 \\ 0 & 2 \end{matrix} \qquad \alpha_2 = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \qquad \beta_2 = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

Then straightforward application of the formulae of the preceeding section yields

$$P = \begin{matrix} 1 & 0 \\ -1 & 1 \end{matrix} \qquad Q = \begin{matrix} -1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{matrix}$$

and the noninteracting controller is shown in Fig. 2b.

The requirement (3.11) may at first seem to be too restrictive. However, all it implies is that the plant cannot be decoupled into sub-systems of order less than k. Hence, if the plant has k state variables in each of its forward paths (see e.q. Fig. 2a), it is then impossible to decouple it into sub-systems of order less than k. Consequently, such plants will necessarily satisfy (3.11). Hence equation (3.11) represents a very realistic requirement without which non-interaction could not be accomplished without increasing the order of the system by means of dynamic equalizers (i.e., without burdening the system with additional state variables).

A byproduct to (3.13) is that each transfer function of the non-interacting sub-systems will have constant numerator. Nevertheless, one still has complete freedom to choose the α's. Hence one still has the complete freedom of choosing the bandwidth of each sub-system (at the expense of higher gain requirements in each feedback loop for an increase in bandwidth) and the location of closed-loop poles. Thus, for example, one can also minimize the settling time (by choosing α's such that all poles of $h^{*}_{ii}(s)$ are equidistant from the jω-axis) or meet other conventional design specifications.

In case where the matrices CA^iB are singular for the first $i = 0, 1, 2, \ldots k-2$ but are not all identically equal to zero, one may (or may not be able to) choose α_i such that

$$\begin{aligned} &\alpha_o \, CB = 0 \\ &\alpha_1 CB + \alpha_o \, CAB \; \alpha \; 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &\alpha_{k-2} \, CB + \alpha_{k-3} \, CAB + \ldots + \alpha_o \, CA^{k-2}B = 0 \end{aligned} \tag{3.16}$$

In this case the design equations (3.14) and (3.15) become

$$P = [\sum_{i=0}^{k-1} \alpha_i \, C \, A^{k-i-1}B]^{-1} \, \beta_k \tag{3.17}$$

and

$$Q = -[\sum_{i=0}^{k} \alpha_i \, C \, A^{k-i-1}B] \, [\sum_{i=0}^{k} \alpha_i \, C \, A^{k-i}] \tag{3.18}$$

Also, as before, $\beta_i = 0$ for all $i = 0,1,2,\ldots k-1$.

Example 2:

The plant

$$\begin{aligned} \dot{x}_1 &= x_3 \\ \dot{x}_2 &= -x_2 + x_4 + x_5 \\ \dot{x}_3 &= x_4 \\ \dot{x}_4 &= u_1 \\ \dot{x}_5 &= -x_5 + u_2 \end{aligned}$$

yields

$$CB = \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \; ; \quad C \, A \, B = \begin{matrix} 0 & 0 \\ 1 & 1 \end{matrix} \; ; \quad C \, A^2B = \begin{matrix} 1 & 0 \\ -1 & -2 \end{matrix}$$

Note that C B and C A B are singular while $C \, A^2B$ is nonsingular. Hence $k = 3$. Then, from (3.16)

$$\beta_1 = \beta_2 = 0$$

and

$$\alpha_o = \begin{matrix} \alpha_{o11} & 0 \\ 0 & 0 \end{matrix}$$

while α_1, α_2, α_3 and β_3 are arbitrary nonsingular diagonal 2 x 2 matrices. Consequently, for noninteraction (3.17) and (3.18) yield, respectively

$$P = \begin{matrix} \dfrac{\beta_{311}}{\alpha_{o11}} & 0 \\ -\dfrac{\beta_{3N}}{\alpha_{o11}} & \dfrac{\beta_{322}}{\alpha_{122}} \end{matrix}$$

and

$$Q = \begin{bmatrix} -\frac{\alpha_{311}}{\alpha_{011}} & 0 & -\frac{\alpha_{211}}{\alpha_{011}} & -\frac{\alpha_{111}}{\alpha_{011}} & 0 \\ \frac{\alpha_{311}}{\alpha_{011}} & -1+\frac{\alpha_{222}}{\alpha_{122}}-\frac{\alpha_{322}}{\alpha_{122}} & \frac{\alpha_{211}}{\alpha_{011}} & 1+\frac{\alpha_{111}}{\alpha_{011}}-\frac{\alpha_{222}}{\alpha_{122}} & 2-\frac{\alpha_{222}}{\alpha_{122}} \end{bmatrix}$$

By direct substitution one can immediately verify that the above controller matrices decoupled the plant into the noninteracting sub-systems

$$\alpha_{011}\, y_1''' + \alpha_{111}\, y_1'' + \alpha_{211}\, y_1' + \alpha_{311}\, y_1 = \beta_{311}\, v_1$$

and

$$\alpha_{122}\, y_2'' + \alpha_{222}\, y_2' + \alpha_{322}\, y_2 = \beta_{322}\, v_2$$

Once again it may be worthwhile to give a physical interpretation of the conditions imposed by (3.16). Since the matrices $CB, \ldots CA^{k-2}B$ are singular (3.16) can be satisfied by non-zero diagonal matrices α_i. Some of the diagonal elements in the α matrices will, however, be zero. This means that the transfer matrix $H^*(s)$ of the noninteracting system will have some of its input-output transfer functions of order less than k. Hence (3.16) does not represent a restriction at all. It simply means that the design procedure developed in this paper will yield non-interacting sub-systems of lowest possible order.

IV. SOME COMMENTS ON STABILITY

A drawback which is common to both the conventional design procedures for multivariable systems and the design by means of state variable feedback is the possibility of hidden instability despite the fact that all the diagonal elements of $H^*(s)$ in (3.8) represent asymptotically stable transfer functions. An interesting example of such hidden instability is contained in [5]. Actually a non-interacting system may contain more state variables than the combined order of all its transfer functions. Hence stability conclusion cannot be made on the basis of transfer functions alone. Instead, one has to check (by means of Routh's test or any other equivalent procedure) that all the eigenvalues of the polynomial

$$\det\left| A + BQ - sI \right|$$

have negative real parts.

If the non-interacting system obtained by the state variable feedback turns out to be unstable one may still be able to obtain a stable non-interacting system by means of a combination of state variable feedback and an equalizer network in the forward path. To do this one augments the differential equations of the plant (2.1) and (2.2) by the equation of the equalizer (2.8) and (2.9). Unfortunately, there is no obvious way to choosing the equalizer such that the number of additional state variables introduced by the equalizer is a minimum. Nor is it possible to predict a priori that a particular choice of additional state variables will definitely eliminate the hidden instability from the non-interacting system. Hence both the phenomenon of hidden instability and the associated problem of its elimination needs further research.

V. CONCLUSIONS

The procedure derived in the paper enables one to design noninteracting controllers for linear plants with cross coupling effects. The noninteraction is accomplished by state variable feedback without any dynamic equalizer networks, thus there is no increase in the order of the overall system above that of the original plant. There are several problems, however, which still have not been answered. First, if the number of system eigenvalues specified by (3.7) is less than the number of plant eigenvalues (i.e., the number of state variables x_i) one may find that the noninteracting system may be unstable[6] even if all the individual transfer functions appear to be asymptotically stable. No general stabilization procedures are presently available to handle such cases without destroying noninteraction between input-output pairs.

In the cases where the present procedures fail it is not known if it may still be possible to design non-interacting controller to yield specific overall transfer functions (i.e., with α's no longer arbitrary).

Finally, in cases where state variable feedback alone cannot accomplish noninteraction, one would like to utilize stable variable feedback in such a manner that noninteraction would require equalizers with minimum number of additional variables.

REFERENCES

1. H. Freeman, "A Synthesis Method for Multipole Control Systems", Transactions AIEE, Vol. 76, Part 2, pp. 28-31, March 1957.
2. R.J. Kavanagh, "Multivariable Control System Synthesis",Transactions AIEE, Vol. 77, Part 2, pp. 425-429, November 1958.
3. I.M. Horowitz, "Synthesis of Linear Multivariable Feedback Control Systems", IRE Trans., Vol. AC-5, No. 2, pp. 94-105, June 1960.
4. J.G. Truxal, Automatic Feedback Control System Synthesis, McGraw-Hill Book Co., New York, N.Y., 1955, Chapter 5.
5. B.S. Morgan, Jr., "The Synthesis of Linear Multivariable Systems by State Variable Feedback", Proc. 1964 JACC, Stanford, Cal., June 1964, pp. 468-472.
6. E.G. Gilbert,"Controllability and Observability in Multivariable Control Systems", SIAM Journal of Control, Ser.A,Vol.2,No.1,pp.128-151, 1963.

LIST OF FIGURES

Fig. 1: Block Diagram Representation of Linear Multivariable Control System Utilizing State Variable Feedback

Fig. 2a: Plant of Example 1

Fig. 2b: Noninteracting Controller of the plant of Example 1

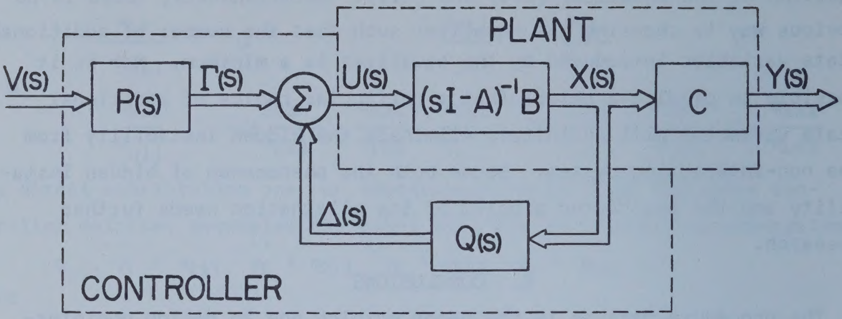

PLANT
V(S)
P(S)
Γ(S)
Σ
U(S)
(sI-A)⁻¹B
X(S)
C
Y(S)
Δ(S)
Q(S)
CONTROLLER

Fig. 1

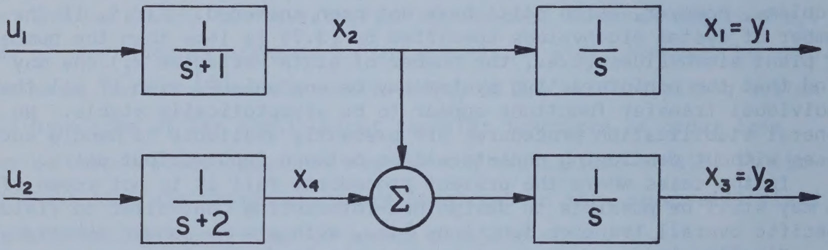

u₁
1/(s+1)
x₂
1/s
x₁ = y₁
u₂
1/(s+2)
x₄
Σ
1/s
x₃ = y₂

Fig. 2a

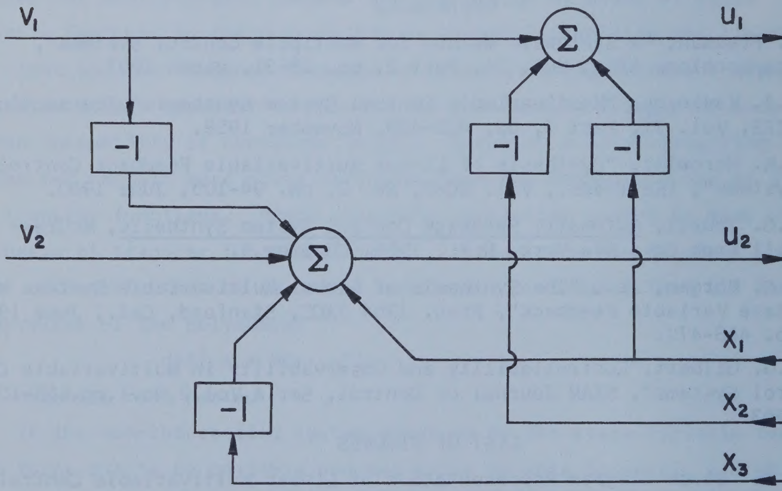

v₁
Σ
u₁
-1
-1
-1
v₂
Σ
u₂
x₁
-1
x₂
x₃

Fig. 2b

EFFECTIVE CONSTRUCTION OF LINEAR STATE-VARIABLE MODELS FROM INPUT/OUTPUT DATA*

B. L. Ho
Stanford University, Stanford, Calif. and
IBM Advanced Development Laboratory,
San Jose, California

R. E. Kalman
Stanford University
Stanford, California

Abstract A new algorithm is given for the effective construction of minimal realizations of a linear, finite-dimensional dynamical system from its external description, which is displayed via the Markov parameters. The algorithm can also be used to prove various abstract results in realization theory; we give a proof of the well-known fact that any two minimal realizations are isomorphic.

I. INTRODUCTION

The basic task of engineering is to translate performance specifications into a blueprint from which a system can be built. We shall be concerned here with a very special but important idealization of this problem which is technically known as the problem of realization [1, 2, 3].

Suppose we are given an external description of a linear dynamical system, such as an impulse-response matrix or a frequency-response matrix. This data may be a performance specification or it may be the result of experimental measurements on a system. We wish to realize the external description via an equivalent internal description, namely a set of state-variable equations. If we are interested in synthesis, these equations are the first step toward producing the blueprint. If we are interested in analyzing experimental data, then these equations are an efficient and useful model with which we can proceed to further analysis and optimization.

The problem of realization for linear, stationary systems was first stated by Elmer Gilbert [4], who gave an algorithm for computing

---

* This research was supported in part by the National Aeronautics and Space Administration under NASA Grant NGR 05-020-073.

The paper is based in part on a forthcoming dissertation of B. L. Ho at Stanford University.

the map: transfer function matrix state-variable differential equations.[*] A second algorithm for the same problem was given at the same time by Kalman [1], using the theory of controllability and observability. Both algorithms require linear-algebra type computations. The deeper theory of the problem consists of a generalization of the classical theory of elementary divisors [2], which is today best expressed in the language of modules [3]. This theory leads to a third algorithm, namely the well-known one which exhibits the canonical form under equivalence of a rectangular polynomial matrix [5, Volume I, Chapter VI].

We present here an entirely different method for computating a realization, which was evolved from the study of the so-called Markov parameters, well known in linear stability theory [6]. The following claims can be made for the new method:

(i) Only very simple numerical operations are needed. In fact, this is probably the simplest method for computing a realization that will ever be found.

(ii) Theoretical properties of realizations can be easily proved via this method. For instance, there is a very simple formula (Proposition 2) for the dimension of a minimal realization.

(iii) Every realization of a given input/output relation can be obtained with this method.

(iv) Given experimental input/output data, there is no need to fix a priori the dimension (= number of state variables) of the system. If the dimension is finite, the method will give an exact realization with the right dimension.

The main limitation of the method is that it is of course restricted to linear systems. Indeed, we know almost nothing today about the nonlinear realization problem.

Our method is so much simpler than its predecessors that we expect to see many new applications. There are also many well-known problems where conventional techniques of data analysis have been cumbersome and expensive; here the new method could bring about major improvements. The following areas seem to us to be especially promising:

---

[*] The scalar case of this problem is quite elementary and has long been known. It is not widely appreciated that the general case is difficult as well as important.

(a) Experimental determination of process dynamics.

(b) Study of the human operator in a control loop.

(c) Biological model building.

(d) Conversion of time-domain data to the frequency-domain and vice versa.

It is only fair to warn the reader that our present results relate only to the noise-free case. We are well aware that the noisy case is the really important one. Research on this aspect of the problem is continuing. Let us point out also that according to research experience [7] major advances in statistical system theory depend mainly on a better understanding of linear deterministic systems. This paper is directed to the latter problem.

2. LINEAR DYNAMICAL SYSTEMS

As is well known (see especially [3]), a finite-dimensional, discrete-time, linear, constant dynamical system has an internal description given by the state-variable (difference) equations

$$(2.1_s) \qquad x(t+1) = \Phi x(t) + Gu(t),\ y(t) = Hx(t)$$

where t = integer and x, u, y are real vectors. The dimension of x is the dimension of the system; u and y have dimension m and p respectively.

Similarly, the internal description of a finite-dimensional, continuous-time linear, constant dynamical system is given by the state-variable (differential) equations

$$(2.2_s) \qquad dx/dt = Fx + Gu(t),\ y(t) = H(t)x(t).$$

The external description of a linear system may be stated either in the time domain or in the transform domain. See [2].

For the system (2.1), the time-domain description is given by the (pulse-response) function

$$(2.1_t) \qquad t \rightarrow Y_t = H\Phi^t G \quad (t = \text{integer} \geq 0);$$

while the transform domain-description is given by the (z-transform) transfer function

$$(2.1_{tr}) \qquad z \to T(z) = H(zI - \Phi)^{-1}G \quad (z = \text{complex number}).$$

For the system (2.2), we have analogously the impulse-response function

$$(2.1_t) \qquad t \to W(t) = He^{Ft}G \quad (t = \text{real number} \geq 0)$$

and the (Laplace-transform) transfer function

$$(2.2_{tr}) \qquad s \to U(s) = H(sI - F)^{-1}G \quad (s = \text{complex number}).$$

The problem of realization is then the following: Given the functions Y or T $[W$ or $U]$ construct constant matrices $[\Phi, G, H]$ [or F, G, H] such that the identities given above hold.

It is implicitly assumed that we are interested only in finite-dimensional realizations, i.e., where F is an $n \times n$ matrix. We make an important preliminary observation:

PROPOSITION 0. The problem stated above is equivalent in all four cases to the following: Given a sequence of $p \times m$ constant matrices $\{Y_k\}_{0 \leq k < \infty}$, find a triple $\{F, G, H\}$ of constant matrices such that

$$(2.3) \qquad Y_k = HF^kG, \quad k = 0, 1, \ldots.$$

Proof. (i) Relation (2.3) is obviously the same as the first realization problem if we set $\Phi = F$ and $t = k$.

(ii) Given the function T, a realization exists only if T is a meromorphic function which is 0 at ∞. Hence in a suitable neighborhood of ∞ T has a power-series representation

$$T(z) = \sum_{k=1}^{\infty} Y_{k-1}z^{-k}.$$

It is clear that $T(z) = H(zI - \Phi)^{-1}G$ if and only if (2.3) holds.

(iii) If W is given, a realization exists only if W is a real analytic function. Hence W has the power series

$$W(t) = \sum_{k=0}^{\infty} Y_k t^k$$

valid for all t; a realization exists if and only if (2.3) holds.

(iv) We have a realization only if U is a meromorphic function and $U(\infty) = 0$, just as in case (ii). For the same reasons, (2.3) is necessary and sufficient for a realization.

In short, for different reasons in each case, an equivalent external description of a finite-dimensional, linear, constant dynamical system (be it discrete-time or continuous-time) is given by the sequence $\{Y_k\}$. So the realization problem is expressible concisely by (2.3). In view of certain classical results [5, Volume II, last chapter; 6], we call $\{Y_k\}$ the Markov parameters of the system.

We say the realization is minimal if the dimension of F satisfying (2.3) is as small as possible [1-3]. The dimension of the minimal realizations is denoted by n_o. If needed, we write $F = F_{min}$ to indicate the fact that F belongs to a minimal realization.

An important fact in realization theory [1-3] is the theorem that all minimal realizations are equivalent modulo the choice of basis used to represent the abstract vector x. This can be proved directly and easily by the present method (Proposition 3).

3. CONSTRUCTION OF A MINIMAL REALIZATION

We shall now describe the new algorithm for the construction of a minimal realization. We assume implicitly that we know the degree r of an annihilating polynomial of F_{min}. For instance, if we are given a transfer-function matrix in rational form, then $r \geq q$, the degree of the least common denominator (= degree of minimal polynomial of F_{min} [2]). If the input/output description is given as experimental data, a special procedure is necessary to determine r; this is discussed in the next section.

PROPOSITION 1. The sequence $\{Y_k\}$ has a finite-dimensional realization if and only if there is an integer r and constants $\alpha_o, \alpha_1, \ldots, \alpha_{r-1}$ such that

(3.1) $$Y_{r+j} = \sum_{i=1}^{r} \alpha_i Y_{r-i+j} \qquad \text{for all} \quad j \geqq 0.$$

The algorithm begins by forming the $r \times r$ block matrix ("generalized Hankel matrix") built out of the Markov parameters

$$S_r = \begin{bmatrix} Y_o & Y_1 & \cdots & Y_{r-1} \\ Y_1 & Y_2 & \cdots & Y_r \\ \vdots & \vdots & & \\ Y_{r-1} & Y_r & \cdots & Y_{2r-2} \end{bmatrix} = [Y_{i+j-2}]$$

Similarly, we write S_r for the $r \times r$ block matrix formed out of the translated sequence $\{Y_{k+\ell}\}$ that is,

(3.2) $$S_r = [Y_{i+j+\ell-2}].$$

PROPOSITION 2. *If* Y_k *has a finite-dimensional realization, then*

(3.3) $$n_o = \text{rank } S_r$$

THEOREM 1. *The following algorithm provides a minimal realization of* $\{Y_k\}$ *whenever a finite-dimensional realization exists:*

(i) *Form* S_r, *where* r *satisfies* (3.1).

(ii) *Find nonsingular matrices* P, Q *(which may be taken to be low and upper triangular) such that*

(3.4) $$PS_rQ = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix} = J$$

(I_s is an $s \times s$ unit matrix, $s = \text{rank } S_r$, and J is idempotent. Note that this is a standard computing problem in linear algebra.)

(iii) *Let* E_p *be the block matrix* $[I_p \quad 0_p \quad \cdots \quad 0_p]$ *and denote by* ulh *the operator which picks out the upper left-hand* *block in block matrices such as* (3.4). *Then*

$$F = \mathrm{ulh}\,[JP(\tau S_r)QJ]$$
$$(3.5) \qquad G = \mathrm{ulh}\,[JPS_r E'_m]$$
$$H = \mathrm{ulh}\,[E_p S_r QJ]$$

is a realization of $\{Y_k\}$. (Proposition 2 shows that this realization is minimal.)

PROPOSITION 3. *Any two minimal realizations* $\{F_i, G_i, H_i\}$, $i = 1, 2$, *of the same* $\{Y_k\}$ *are isomorphic; that is, there exists a nonsingular matrix* T *such that*

$$(3.6) \qquad F_2 = TF_1T^{-1}, \quad G_2 = TG_1, \quad H_2 = H_1T^{-1};$$

in fact, T *is given explicitly by* $(V_2V_2')^{-1}V_2V_1'$ (notation explained below).

PROPOSITION 4. *Every minimal realization may be obtained by suitably choosing* P *and* Q *in the algorithm described under Theorem 1.*

We shall now present the main ideas of the proofs. A precise demonstation may be found in [8].

For simplicity, we assume that r is suitably fixed and drop it as a subscript.

Necessity of (3.1). Let $\psi(z) = \beta_0 z^q + \beta_1 z^{q-1} + \ldots + \beta_q$, $\beta_0 = 1$, be the minimal polynomial of some $F_{\min}$. Then, if $Y_k = HF^kG$,

$$0 = HF^j\psi(F)G = \sum_{i=0}^{q} \beta_i HF^{q-i+j}G$$

so that (3.1) holds with $r = q$ and $\alpha_i = -\beta_i$. In fact, q just defined is the minimum value of r for which (3.1) holds.

Sufficiency of (3.1) follows from Theorem 1.

Proof of Proposition 2. Take a minimal realization $\{F, G, H\}$. Write $S = V'W$ (' = transpose), where

$$V = [H' \quad F'H' \quad \ldots \quad (F')^{r-1}H']$$

$$W = [G \quad FG \quad \ldots \quad F^{r-1}G]$$

Since F is $n_o \times n_o$ it is clear that rank $V \leqq n_o$ and rank $W \leqq n_o$. Hence also rank $S \leqq n_o$. If rank $S < n_o$, then by Theorem 1 there exists a realization of dimension $< n_o$. This contradicts the minimality of n_o.

Proof of Theorem 1. Form the "block companion matrix"

$$M = \begin{bmatrix} 0_p & I_p & 0_p & \ldots & 0_p \\ 0_p & 0_p & I_p & \ldots & 0_p \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0_p & 0_p & 0_p & \ldots & 0_p \\ \alpha_r I_p & \alpha_{r-1} I_p & \alpha_{r-2} I_p & \ldots & \alpha_1 I_p \end{bmatrix}$$

Observe that, by (3.1),

$$\tau^k S = M^k S, \quad k = 0, 1, \ldots \quad . \tag{3.7}$$

Similarly, there is an $r \times r$ block companion matrix N (given by the transpose of the block pattern of M but with $m \times m$ blocks) such that

$$\tau^k S = SN^k, \quad k = 0, 1, \ldots \quad . \tag{3.8}$$

It is easy to verify that

$$Y_k = E_p(\tau^k S)E'_m = E_p M^k SE'_m, \qquad k \geqq 0, \tag{3.9}$$

so that $\{M, SE'_m, E_p\}$ is a realization. In general, it is nonminimal. It must be therefore reduced so that its dimension becomes equal to rank S. How this is done is difficult to motivate but easy to demonstrate.

Write $S^{\#} = QJP$. Then $S^{\#}$ is a pseudo-inverse of S, that is, $S = SS^{\#}S$. With the aid of this identity and the fact that $M^k S = SN^k$ (see (3.7-8)), we rearrange (3.9) successively as follows

$$\begin{aligned} Y_k &= E_p M^k SE'_m \\ &= E_p SS^{\#} M^k SS^{\#} SE'_m \\ &= (E_p SQJ)(JPM^k SQJ)(JPSE'_m) \\ &= (E_p SQJ)(JPMSQJ)^k (JPSE'_m). \end{aligned}$$

Hence by (3.7) $\{JP(\tau S)QJ,\ JPSE'_m,\ E_p SQJ\}$ is a realization. The 0 blocks in these matrices may be dropped without affecting the fundamental relation (2.3).

Proof of Proposition 3. Each of the four matrices $\{V_i, W_i\}$, $i = 1, 2$, associated with the two minimal realizations have n_o rows, and each of them has rank n_o. So $V_i V'_i$ and $W_i W'_i$ are all nonsingular. By the definition of a realization $V'_1 W_1 = S = V'_2 W_2$. Multiplying on the left by V_2 we get $TW_1 = W_2$, with $T = (V_2 V'_2)^{-1} V_2 V'_1$. T is nonsingular because if $U = W_1 W'_2 (W_2 W'_2)^{-1}$, then $TU = I$. Thus $V'_1 U = V'_1 T^{-1} = V'_2$. But $\tau S = V'_1 F_1 W_1 = V'_2 T F_1 T^{-1} W_2 = V'_2 F_2 W_2$, which implies that $F_2 = T F_1 T^{-1}$. The remaining relations may be easily verified.

Proof of Proposition 4. We must exhibit a pair P_2, Q_2 such that a given minimal realization $\{F_2, G_2, H_2\}$ is expressed by formulas (3.5). First we take any $P = P_1$ and $Q = Q_1$ for which (3.4) holds and construct the corresponding realization $\{F_1, G_1, H_1\}$. Then we let $T = (V_2 V'_2)^{-1} V_2 V'_1$ and

$$P_2 = \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} P_1, \qquad Q_2 = Q_1 \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix}$$

Relation (3.4) is preserved. By Proposition 3 we see that the realization corresponding to P_2, Q_2 is $\{F_2, G_2, H_2\}$.

Two points deserve special emphasis:

(i) The intermediate step of constructing the block canonical matrix M is a theoretical trick. It can be carried out only if the coefficients of an annihilating polynomial are known. As a rule, they cannot be determined directly from $\{Y_k\}$. If the initial data is given in rational transfer-function form, then the desired polynomial is simply the least common denominator.

(ii) The formulas for the minimal realization no longer involve M because of the relation $MS = \tau S$. (τS is known of course from $\{Y_k\}$.) Thus the only piece of information required in addition to a suitable number of Y_k's is the integer r. This is an important aspect of the new algorithm. Trying to find the constants $\alpha_o, \ldots, \alpha_{r-1}$, constructing M, and then reducing the realization would involve much labor which is avoided by applying formulas (3.5) directly. In fact, it is not even necessary to have a recurrence relation such as (3.1) (see Theorem 2). We have here a strong plausibility argument that the algorithm given by Theorem 1 is maximally efficient.

4. DETERMINATION OF r

The procedure described in the preceding section requires only one abstract assumption, namely knowledge of the integer r in the recurrence relation (3.1). How can r be determined?

If $\{Y_k\}$ has no finite-dimensional realization, then we are faced with an approximation problem. We have no significant results as yet in this area.

If $\{Y_k\}$ has a finite-dimensional realization, the following results provide a method for effectively choosing r:

THEOREM 2. Suppose that rank S_r = rank S_{r+1} for some integer r. Then if $\{F, G, H\}$ are given by (3.5) the fundamental relation (2.3) is satisfied for all $0 \leq k \leq 2r - 2$, that is, for every element of the matrix S_r.

COROLLARY. If rank S_t = constant for all $t \geqq t_o$, then (3.5) with $r = t_o$ provides a realization of $\{Y_k\}$.

The proofs of these results require subtle modifications of the proof of Theorem 1 and may be found in [8].

We have now a practical procedure for computing a realization which does not require looking at infinitely many Y_k's: We examine in succession the rank of $S_1, S_2, \ldots$ until rank S_r = rank S_{r+1}.

Then we compute $\{F, G, H\}$ by (3.5) and set $\hat{Y}_k = HF^kG$. The difference sequence formed by the $\tilde{Y}_k = Y_k - \hat{Y}_k$ will have 0's for at least its first $2r-1$ elements. In view of formula (2.1_{tr}), if the $\{\tilde{Y}_k\}$ are not identically zero and have a finite-dimensional realization, they correspond to a transfer function matrix which has s^p as a factor in its least common denominator, with $p \geq 2r-2$. This observation allows us to dispense with the 0 terms in the $\{\tilde{Y}_k\}$ sequence and consider only sequences in which $Y_o \neq 0$. By repeating this procedure, we eventually obtain a realization *if* a finite-dimensional realization exists at all. Note that it may be possible to extract a finite-dimensional part of a sequence which has no finite-dimensional realization.

With respect to the corollary, it is interesting to observe that if r is the least number such that rank S_t = constant for $t > r$, then r may be strictly less than the degree q of the minimal polynomial of F. For example, take $F = n \times n$ diagonal matrix, with distinct nonzero diagonal elements, $G = I_n$, and $H = I_n$. Then $\{F, G, H\}$ realizes $\{F^k\}$ minimally. We see that rank $S_1 = n$ and hence $r = 1$, while $q = n$.

REFERENCES

[1] R. E. Kalman, "Mathematical description of linear dynamical systems", J. Control (SIAM), 1 (1963) 152-192.

[2] R. E. Kalman, "Irreducible realizations and the degree of a rational matrix", SIAM J., 13 (1965) 520-544.

[3] R. E. Kalman, "Algebraic structure of linear dynamical systems. I. The module of Σ", Proc. Nat. Acad. Sci. (USA), 53 (December 1965).

[4] E. G. Gilbert, "Controllability and observability in multivariable systems", J. Control (SIAM), 1 (1963) 128-151.

[5] F. R. Gantmakher, THE THEORY OF MATRICES (book), Chelsea, 1959.

[6] F. Jarominek, First and Second IFAC Congress, 1960 and 1963.

[7] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory", J. Basic Engr. (Trans. ASME), 83 D (1961) 95-108.

[8] B. L. Ho and R. E. Kalman, "Markov parameters, the moment problem, and minimal realizations", J. Control (SIAM), 4 (1965) (to appear).

NONLINEAR TIME VARYING FEEDBACK SYSTEMS -- CONDITIONS FOR L_∞-BOUNDEDNESS DERIVED USING CONIC OPERATORS ON EXPONENTIALLY WEIGHTED SPACES.

G. ZAMES
Dept. of Electrical Engineering, MIT, Cambridge, Mass.

1. ABSTRACT AND INTRODUCTION

In the design of input-output feedback systems such as feedback amplifiers, phase-locked loops, etc., it is required to predict, and if possible prevent, instabilities in which bounded inputs produce unbounded outputs. The absence of such instabilities is subsumed under the mathematical condition of L_∞-boundedness. This abbreviated paper has two main results: Theorem 1 is a functional L_∞-boundedness condition, general in the sense that no restrictions are placed on linearity or time invariance of the open loop feedback elements (Fig. 1a) or on dimensionality or continuity of inputs; input-output relations of the loop elements are instead confined to certain cone-like regions in function space, having prescribed center and radius parameters, reminiscent of contours of the Nyquist-Nichols theory. Theorem 2 is a frequency response, circle condition for loops (Fig. 1b) consisting of (i) a linear time invariant element, (ii) a Lipschitz sector nonlinearity, (iii) a time-varying gain, and (iv) a linear "R.C." multiplier having poles and zeroes alternating on the negative real axis.

1.1 Background

Loops containing one nonlinearity have been the subject of more treatises than can properly be acknowledged in this limited space. Especially relevant are: Popov's [1] well known condition; circle criteria obtained by Lyapunov methods by Narendra and Goldwyn [2]; circle criteria obtained by functional means by Kudrewicz [3] and by the author [4a]; the method of factoring an open loop into two positive operators, exploited by the author in a functional setting, and by Brockett, Willems and Forys in a Lyapunov setting, which has led to stability theorems involving R.C. multipliers - monotone-nonlinearities [4a and 5a] and a time varying gain [5b].

2. FORMULATION

Our first task is to define a suitable space X_e of time functions; an input-output relation will then be taken to be a mapping of X_e into itself, and feedback equations will be written. Such elementary notions of functional analysis as those of "linear space," "norm", and "inner product", will be used freely.

2.1 Spaces

It can usually be deduced, from coarse considerations of open loop delay, that a feedback loop subjected to a locally bounded input on $[0,\infty)$ has a unique, locally bounded output--that there are no finite "escape times." We wish to make this fact the starting point of our analysis, and turn to

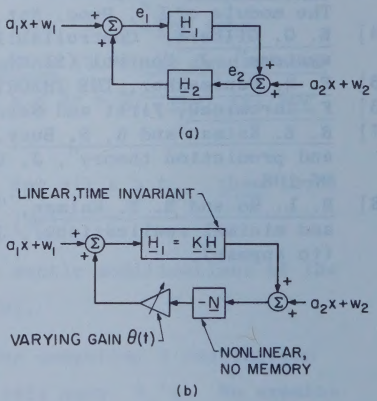


Fig. 1 Feedback Loops

the finer problem of showing that L_∞ inputs produce L_∞ outputs. Hence our basic space X_e consists of functions bounded locally, but not necessarily bounded on the whole interval. The following are definitions:

T - is a fixed interval of the type $[t_0,\infty)$ or $(-\infty,\infty)$.

V - is some fixed, linear, innerproduct, normed space; the innerproduct of x and y in V is $x \cdot y$; the norm of x in V is $|x| = \sqrt{x \cdot x}$

L_2 - is the space of square integrable, V-valued functions on T; it is assumed that L_2 is a linear, innerproduct, normed space; the inner product of x and y in L_2 is

$$<x, y> = \int_T x(t) \cdot y(t)\, dt$$

and the norm of x is $\|x\| = \sqrt{<x,x>}$. We shall not usually distinguish between functions equal almost everywhere.

L_∞ -is the space of measurable, bounded, V-valued, functions on T. Assume that L_∞ is a linear, normed space, and the norm of $x \in L_\infty$ is $\|x\|_\infty = \operatorname{lub}_{t \in T} |x(t)|$.

x_t - *Notation.* If x is a function on T and $t \in T$ let x_t be the *truncated function* on T defined by: $x_t(\tau) = x(\tau)$ for $\tau < t$ and $x_t(\tau) = 0$ elsewhere.

$L_{\infty e}$ -is the space of those functions x on T for which $x_t \in L_\infty$ for all $t \in T$. Observe that $L_{\infty e}$ is linear and partially normed by $\| \;\|_\infty$.

X_e -is a fixed, linear, subspace of $L_{\infty e}$.

2.2 Input-Output Relations

It is convenient to divorce the problem of boundedness from that of existence-uniqueness, since the two can often be attacked by disparate methods. The existence-uniqueness problem will be ignored by representing a system by a "relation" (see, for example, J. L. Kelley, "General Topology" p. 6) which, by contrast with an operator, can be multivalued and undefined on parts of X_e. In effect we concentrate on "such solutions as do exist."

A (input-output or output-input) relation H is defined to be any set of ordered pairs (x,y) of elements of X_e; the set of first elements x is the *domain* of H denoted $\mathrm{Do}(H)$; the set of second elements y is the *range* of H denoted $\mathrm{Ra}(H)$. If $x \in \mathrm{Do}(H)$ then Hx denotes any one of the elements of $\mathrm{Ra}(H)$ with the property that (x, Hx) belongs to H.

Convention. An expression containing Hx free of quantifiers "for all", "there exists", etc., is assumed to hold *for all* Hx in $\mathrm{Ra}(H)$.

Notation. $Hx(t)$ denotes the value assumed at time $t \in T$ by (some function which almost everywhere equals) the function Hx.

An *operator* is any single valued relation whose domain is the whole of X_e.

Let $\mathcal{R}$ be the set of all relations and let $\mathcal{R}_0$ be the subset of $\mathcal{R}$ consisting of those $H \in \mathcal{R}$ for which 0 is in $\mathrm{Do}(H)$ and $H(0) = 0$.

Notation: I is the identity operator on X_e. If a and b are real constants and $K, H \in \mathcal{R}$ then $aH + bK$ denotes the relation whose domain is $\mathrm{Do}(H) \cap \mathrm{Do}(K)$ and for which $(aH + bK)x = aHx + bKx$.

2.3 Boundedness; Memory Constraint

In order to exclude the possibility that bounded inputs produce unbounded -- or even bounded but arbitrarily large -- outputs, we adopt the following definition: $H \in \mathcal{R}$ is *L_∞-bounded* if H takes L_∞-bounded subsets of $\mathrm{Do}(H)$ into L_∞-bounded subsets of $\mathrm{Ra}(H)$; a subset Y of X_e is L_∞-bounded if $\|y\|_\infty <$ const. for $y \in Y$.

We shall need a smoothing condition of the following type: $\underline{H} \in \mathcal{R}_o$ has decaying L_1-memory if there exists an absolutely integrable, nonnegative, nonincreasing, real valued function h on $[0,\infty)$ for which, (t_o being the starting point of T,)

$$|\underline{H}e(t)|^2 \le \int_{t_o}^{t} |e(\tau)|^2 \, h(t-\tau)\, d\tau \qquad (e \in Do(\underline{H});\ t \in T) \qquad (1)$$

2.4 Conic and Positive Relations

Our theory will confine open-loop input-output pairs to regions of an exponentially weighted function space; it will appear that for memoryless relations these regions take the form of conic sectors in the instantaneous input-output plane; hence "conic" relations. The notation $\exp^{\sigma}$ will denote the exponential function on T whose value at $t \in T$ is $\exp(\sigma t)$. Observe that if $x \in X_e$ then $x_t \exp^{\sigma} \in L_2$ for $\sigma > 0$, $t \ge 0$.

A relation $\underline{H} \in \mathcal{R}_o$ is conic (c,r), where $r \ge 0$ and c are real constants, if for some $\sigma < 0$

$$\|(\underline{H}x - cx)_t \exp^{-\sigma}\| \le r\,\|x_t \exp^{-\sigma}\| \qquad (x \in Do(\underline{H});\ t \in T) \qquad (2)$$

If the inequality in (2) is replaced by $\ge$ then $\underline{H}$ will be said to be -conic (c,r). (2) implies that the weighted, truncated, output $(\underline{H}x)_t \exp^{\sigma}$ lies in a sphere in L_2, with center and radius proportional to the input. c and r will be referred to as "center" and "radius" parameters respectively. Note the resemblance between (2) with $\exp^{-\sigma}$ deleted, and the equation $|y-cx| \le r|x|$ which describes a conic sector in the two-dimensional (x,y) plane, with center line and spread determined by the parameters c and r respectively.

A relation $\underline{H} \in \mathcal{R}_o$ is positive if for some $\sigma < 0$

$$\langle x_t \exp^{-\sigma}, (\underline{H}x)_t \exp^{-\sigma} \rangle \ge 0 \qquad (x \in X_e;\ t \in T) \qquad (3)$$

A positive relation can be viewed as degenerately conic, with conic sector replaced by a half space.

2.5 Feedback Equations

Consider the loop of Fig. 1a; $\underline{H}_1$ and $\underline{H}_2$ are fixed relations in $\mathcal{R}_o$; x is an input added in at two points to get the two situations of interest in one analysis; a_1 and a_2 are, in the simplest case, real constants -- by setting either a_1 or a_2 to zero one of the inputs can be eliminated; more generally, assume that a_1 is a relation satisfying an inequality of the type $\|a_1 x\|_\infty \le \text{const.}\|x\|_\infty$ and similarly for a_2. To account for effects of initial conditions we include two fixed biases, w_1 and w_2 in $X_e \cap L_\infty$. Equations are:

$$e_1 = w_1 + a_1 x + \underline{H}_2 e_2 \quad (4a) \qquad e_2 = w_2 + a_2 x + \underline{H}_1 e_1 \quad (4b)$$

Let $\underline{E}_1$ in $\mathcal{R}$ be the relation consisting of those pairs (x, e_1) of elements of X_e for which there exist e_2, $\underline{H}_1 e_1$, and $\underline{H}_2 e_2$, all in X_e, satisfying (4). Similarly, let $\underline{E}_2$ in $\mathcal{R}$ consist of (x, e_2).

3. RESULTS FOR LOOPS CONSISTING OF CONIC PAIRS

For equations (4) and the loop of Fig. 1a we seek conditions on $\underline{H}_1$ and $\underline{H}_2$ under which $\underline{E}_1$ and $\underline{E}_2$ are L_∞-bounded. It develops that if $\underline{H}_1$ and $\underline{H}_2$ are conic it is enough for their center and radius parameters to be proportioned as follows:

Theorem 1. Let c, $r \geq 0$ and $\Delta \geq 0$ be constants. Suppose $\underline{H}_1$ is conic $(-c, r-\Delta)$; sufficient conditions for $\underline{E}_1$ and $\underline{E}_2$ to be L_∞-bounded are:

(i) Case 1a. $r^2 < c^2$. $\underline{H}_2$ is -conic $(\gamma, \rho + \delta)$, where $\gamma = c/(r^2 - c^2)$; $\rho = r/(r^2 - c^2)$; $\delta > 0$ is a constant greater than zero if $\Delta = 0$ and zero if $\Delta > 0$.

Case 1b. $r^2 > c^2$. $\underline{H}_2$ is conic $(\gamma, \rho - \delta)$.

Case 2. $r^2 = c^2$. $[2c\underline{H}_2 + (1 - \delta)\underline{I}]$ is positive; in addition, if $\Delta = 0$ then $\underline{H}_2$ is conic (0, const.).

(ii) $\underline{H}_1$ has decaying L_1-memory and $\underline{H}_2$ is L_∞-bounded or conversely.

Proof. Only the situation $\Delta > 0$, $\delta = 0$, will be considered; the alternative situation can be derived from the first by a transformation in which r is replaced by a larger parameter. *Notation.* Let

$$\|x; t, \sigma\| = \|x_t \exp^{-\sigma}\| \qquad <x, y;\ t, \sigma> = <x_t \exp^{-\sigma},\ y_t \exp^{-\sigma}>$$

In the remainder of this proof suppose that $(x, e_1, e_2, \underline{H}_1 e_1, \underline{H}_2 e_2)$ is any quintuplet of elements of X_e satisfying (4). The hypothesis implies, in Case 1b, that for some $\sigma_1 < 0$, $\sigma_2 < 0$,

$$\|\underline{H}_1 e_1 + c e_1;\ t, \sigma_1\| - (r - \Delta)\|e_1;\ t, \sigma_1\| \leq 0 \qquad (t \in T) \tag{5}$$

$$\|\underline{H}_2 e_2 - \gamma e_2;\ t, \sigma_2\| - \rho\|e_2;\ t, \sigma_2\| \leq 0 \qquad (t \in T) \tag{6}$$

It can be shown by integration by parts that if an inequality of type (5) holds for $\sigma_1 < 0$ it holds for all σ in the range $\sigma_1 \leq \sigma < 0$; also (6) holds for $\sigma_2 \leq \sigma < 0$; let $\mu = \max(\sigma_1, \sigma_2)$; hence (5) and (6) hold with σ_1 and σ_2 replaced by μ.

Step 1. It will be shown that there exist positive constants a,b,c,d, for which, for $x \in L_\infty$, $t \in T$,

$$\exp \mu t\ \|e_1; t, \mu\| \leq a + b\|x\|_\infty \quad (7) \qquad \exp \mu t\ \|e_2; t, \mu\| \leq c + d\|x\|_\infty \quad (8)$$

From (6) get

$$<\underline{H}_2 e_2 - (\gamma - \rho)e_2,\ \underline{H}_2 e_2 - (\gamma + \rho)e_2;\ t, \mu> \ \leq 0 \qquad (t \in T)$$

Hence after some manipulation

$$r\|\underline{H}_2 e_2;\ t, \mu\| - \|c\underline{H}_2 e_2 + e_2;\ t, \mu\| \leq 0 \qquad (t \in T) \tag{9}$$

Cases 1a and 2 also give (9). Substituting (4a) for $\underline{H}_2 e_2$ and (4b) for e_2 into (9), and using the triangle inequality, get

$$r\|e_1;\ t, \mu\| \leq \{\|(w_2 - cw_1) + (a_2 - ca_1)x;\ t, \mu\| + \|w_1 + a_1 x;\ t, \mu\|\} + \|\underline{H}_1 e_1 + c e_1;\ t, \mu\| \tag{10}$$

Now for any $y \in L_\infty$ the inequality $\|y; t, \mu\| \leq \|y\|_\infty (\exp -\mu t)/\sqrt{-2\mu}$ holds; but w_1, w_2 and x are in L_∞; recall also that $\|a_1 x\|_\infty \leq$ const. $\|x\|_\infty$; hence the {} bracketed term in (10) is bounded by an expression $\exp -\mu t\ (a_o + b_o\|x\|_\infty)$ where a_o and b_o are positive constants; the other term on the r.h.s. of (10) is bounded using (5); hence

$$\Delta \exp \mu t\ \|e_1; t, \mu\| \leq a_o + b_o\|x\|_\infty \tag{11}$$

Δ being positive; (7) follows. Furthermore, from (4b), the triangle inequality and (5) we obtain

$$\|e_2;\ t, \mu\| \leq \|w_2 + a_2 x;\ t, \mu\| + (r + |c|)\|e_1;\ t, \mu\| \tag{12}$$

From (12), (11), and since $(w_2 + a_2 x) \in L_\infty$, get (8).

Step 2. Suppose $\underline{H}_1$ has decaying memory and $\underline{H}_2$ is L_∞-bounded. (The following argument holds also if $\underline{H}_1$ and $\underline{H}_2$ are interchanged.) It will be shown that

$$\|e_2\|_\infty \leq \text{const.} + \text{const.} \|x\|_\infty \tag{13}$$

so that $\underline{E}_2$ is L_∞-bounded. The L_∞-boundedness of $\underline{E}_1$ is deducible from that of $\underline{E}_2$ and $\underline{H}_2$, and from (4a). Hence Theorem 1 is true.

To prove (13) observe that the memory constraint implies an inequality of type (1). Now $|e_1(\tau)|^2$ can be expressed a.e. as the inverse of a convolution,

$$|e_1(\tau)|^2 = J(\tau) - \frac{1}{2\mu}\frac{dJ}{d\tau} \quad \text{where } J(t) = \int_{t_o}^{t} |e_1(\tau)|^2 \exp 2\mu(t-\tau)\, d\tau \tag{14}$$

Combining (14) and (1) obtain

$$|\underline{H}_1 e_1(t)|^2 \leq \int_{t_o}^{t} h(t-\tau)\, J(\tau)\, d\tau - \frac{1}{2\mu}\int_{t_o}^{t} h(t-\tau)\frac{dJ}{d\tau}\, d\tau \tag{15}$$

$$= \int_{t_o}^{t} h(t-\tau)\, J(\tau)\, d\tau - \frac{h(0)J(t)}{2\mu} + \frac{1}{2\mu}\int_{t_o}^{t} J(\tau)\, d_\tau h(t-\tau) \tag{16}$$

the second integral in (15) having been integrated by parts (recall, h is monotone). Now $J(\tau) = \exp 2\mu\tau \|e_1; \tau, \mu\|^2$; hence (7) gives $J(\tau) \leq (a' + b'\|x\|_\infty)^2$; now (16) is bounded,

$$|\underline{H}_1 e_1(t)|^2 \leq (a' + b'\|x\|_\infty)^2 \quad \int_0^\infty |h(\tau)|d\tau + \frac{h(0)}{2\mu} + \frac{1}{|2\mu|}\int_{-\infty}^{0} d_\tau h(-\tau)$$

$$= (a' + b'\|x\|_\infty)^2 \cdot \{\text{const.}\} \tag{17}$$

Finally (13) is obtained from (17) and (4b). QED.

Comments. L_∞-boundedness has been deduced from rather gross assumptions on open loop input-output behavior; the open loop need neither be linear nor time invariant.

There is a resemblance between Nyquist's criterion and the following limiting form of Theorem 1: "If $-\underline{H}_1$ equals $k\underline{I}$, a sufficient condition for L_∞-boundedness is for $\underline{H}_1$ to be bounded away from the critical value $-k^{-1}\underline{I}$ by some radius δ, i.e., for $\underline{H}_2$ to be -conic $(-k^{-1}, \delta)$." However, unlike Nyquist's criterion, Theorem 1 requires observations for all inputs in X_e -- not merely the frequency response; moreover, Theorem 1 is only sufficient -- hardly surprising, since detailed behavior has been neglected.

Theorem 1 includes results of the type (see [4a]). "If the open loop gain is less than unity the closed loop is bounded," (obtained by setting $c = 0$, $r = 1$,) and "if the open loop is factorizable into one positive operator $\underline{H}_2$, and one strongly positive, finite gain, operator $\underline{H}_1$, the closed loop is bounded," (obtained by setting $c = r >>$ gain^2 of $\underline{H}_1$).

4. LOOPS CONTAINING ONE LINEAR OPERATOR, ONE MONOTONE NONLINEARITY, AND ONE TIME VARYING GAIN

Turn now to the loop of Fig. 1b in which $\underline{H}_1$ is time invariant, linear, $\underline{N}$ is time invariant, memoryless, and $\underline{\theta}$ is a time varying gain. In exchange for suitable restrictions on $\underline{N}$ and $\underline{\theta}$ we seek regions in which the frequency response $H_1(j\omega)$ can safely lie without fear of instability. Our method is to try to factor the open loop into two conic operators, in accordance with Theorem 1. The simplest result, obtained by choosing $\underline{H}_1$ and $\underline{\theta\, N}$ as the two operators, places $\underline{N}$ in a sector of the $\underline{N}x(t)$ vs. $x(t)$ plane, and $H(j\omega)$ inside or outside a related disk in the complex frequency plane (see [4a], [2]). We seek to improve on this simple result by first splitting the upper part of the loop, $\underline{H}_1$, into two factors $\underline{K}$ and $\underline{H}$, and shifting $\underline{K}$ down and ahead of $\underline{\theta\, N}$. The point here is to find a factor $\underline{K}$ which narrows the cone of confinement of the top more than it widens that of the bottom -- so as to obtain a wider safe region for the frequency response.

4.1 Definitions

Restrictions. The interval T is fixed at $[0,\infty)$. $L_{\infty e}$ is defined as in Sect. 2.1, but its members are real valued. X_e is assumed equal to $L_{\infty e}$. All relations are assumed operators, i.e., single valued and defined for all x in X_e.

Notation. "$\underline{H}$ is confined to the sector $\{a,b\}$" means that $a \le b$ are real constants and for some $\sigma < 0$

$$<(\underline{H}x-ax)_t\exp^{-\sigma},\ (\underline{H}x-bx)_t\exp^{-\sigma}> \ \le\ 0 \qquad (X\epsilon L_{\infty e};\ t\ge 0 \qquad (18)$$

It can be verified that this is another way of saying that $\underline{H}$ is conic $(1/2(b+a),\ 1/2(b-a))$; in certain situations the former notation is more convenient than the latter -- when in forming operator products sector parameters multiply while cone parameters undergo a more complicated transformation.

The class $\mathcal{O}$ consists of all operators mapping $L_{\infty e}$ into itself. The sum $(\underline{K}+\underline{H})$, product by a scalar $c\underline{H}$, and composition product $\underline{K}\,\underline{H}$ of $\underline{K}$ following $\underline{H}$, are defined in $\mathcal{O}$ in the usual way.

The class $\mathcal{L}$ consists of all $\underline{H}\epsilon\, \mathcal{O}$ satisfying

$$\underline{H}\,x(t) = h_\infty x(t) + \int_0^t x(\tau)\, h(t-\tau)\, d\tau \qquad (x\epsilon L_{\infty e};\ t\ge 0 \qquad (19)$$

where (1) h_∞ is a real constant, (ii) $h(t)$ is a real valued function on $[0,\infty)$, and (iii) $h(t)\exp -\mu_o t$ is absolutely integrable on $[0,\infty)$, $\mu_o < 0$ being constant.

The Laplace transform $H(s)$ of $\underline{H}\epsilon\,\mathcal{L}$ is

$$H(s) = \int_0^\infty h(t)\exp(-st)\,dt \qquad (\sigma\ge\mu_o)$$

and converges for $\sigma\ge\mu_o$, where $s=\sigma+j\omega$ denotes a point in the complex plane.

The μ-shifted Nyquist diagram of $H(s)$ is the contour consisting of the points $H(\mu+j\omega)$, (ω being real,) augmented with the point h_∞.

The class $\mathcal{RC}$ consists of those $\underline{K}\epsilon\,\mathcal{L}$ having transforms

$$K(s) = K(0) + \sum_{i=1}^{m} \frac{k_i}{s+p_i} \qquad (0<p_1<p_2\cdots<p_m;\ k_i\geq 0;\ K(0)\geq 0)$$

(K(s) can be realized as the driving point impedance of a resistance-capacitance network -- hence "$\mathcal{RC}$".)

The class $\mathcal{N}$ consists of those $\underline{N}\in\mathcal{O}$ satisfying

$$\underline{N}x(t) = N(x(t)) \qquad (x\in L_{\infty e};\ t\geq 0)$$

where (1) N is a real valued function on the reals, (ii) $N(0)=0$, (iii) N is confined to a sector $\{a,b\}$, that is $a\leq N(x)/x \leq b$ for real $x\neq 0$, (iv) $\underline{N}$ satisfies the monotonicity condition

$$a(x-y)^2 \leq (x-y)[N(x)-N(y)] \leq b(x-y)^2 \qquad (x,y \text{ real}) \qquad (20)$$

Condition (iv) can be replaced by the weaker condition that N(x) is the difference of two nondecreasing parts, $N(x) = N_+(x)-N_-(x)$, confined to respective sectors $\{0,b\}$ and $\{0,-a\}$.

The class $\mathcal{J}$ consists of those $\underline{\theta}\in\mathcal{O}$ satisfying $\underline{\theta}x(t) = \theta(t)\cdot x(t)$ where (i) $\theta(t)$ is a real valued function on $[0,\infty)$, (ii) $\theta(t)>0$ for $t\geq 0$, and (iii) $\underline{\theta}$ is confined to some sector $\{a,b\}$, that is, $a\leq\theta(t)\leq b$ for $t\geq 0$, where $a>0$.

4.2 Lemmas Giving Conic Bounds on Various Operators

In preparation for our main task let us estimate center and radius parameters, (or sector parameters), for the various operators.

Lemma 1. Suppose (i) $\underline{N}\in\mathcal{N}$ is nondecreasing, (ii) $\underline{K}\in\mathcal{RC}$ has no poles more positive than $-p_1$ where $p_1>0$, and (iii) $\underline{\theta}\in\mathcal{J}$ has $\theta(t)\exp -\lambda t$ nonincreasing where $\lambda<p_1$ is constant -- then $\underline{\theta}\,\underline{N}\,\underline{K}$ is positive.

Note: For our purposes the significant aspect of an $\mathcal{RC}$ operator $\underline{H}$ is that its product with a nonlinearity $\underline{N}\in\mathcal{N}$ is positive -- at least if $\underline{N}$ follows $\underline{H}$; in general the product of two positive operators need not be positive. Furthermore, the next Lemma implies that the rule for confining $\underline{N}\,\underline{H}$ in a sector is the same as if $\underline{H}$ had no memory -- giving a smaller sector than would generally be obtained. Finally, observe that the condition on θ implies that the increasing part of θ has a derivative not greater than λ. (The decreasing part of θ need not be differentiable.)

Note. It can be shown by a partial integration that if $\underline{H}$ is positive, (or confined to some given sector), and $\underline{\theta}\in\mathcal{J}$ is nonincreasing -- then $\underline{\theta}\,\underline{H}\,\underline{\theta}^{-1}$ is also positive, (or confined to that sector).

Lemma 1a. Suppose (i) $\underline{K}\in\mathcal{RC}$ has no poles more positive than $-p_1$, (ii) $\underline{N}\in\mathcal{N}$, $\underline{\theta}\in\mathcal{J}$, and $\underline{\theta}\,\underline{N}$ is confined to a sector $\{a, b-\Delta'\}$, where $a\geq 0$, $b>0$, and $\Delta'>0$ are constants, (iii) $\underline{\phi}\in\mathcal{J}$, and $\phi^{-2}\theta\exp -\lambda t$ is nonincreasing, where $\lambda<p_1$ -- then there exists $\Delta>0$ for which $\underline{\phi}^{-1}\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi}$ is confined to the sector $\{aK(\infty),\ bK(-\lambda)-\Delta\}$.

Lemma 1b. If the hypothesis of Lemma 1a holds, except that a is replaced by $a'+\Delta'$ where $a'<0$ -- then $\underline{\phi}^{-1}\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi}$ is in $\{a'K(-\lambda)+\Delta,\ bK(-\lambda)-\Delta\}$.

Lemma 2. Case 1. Suppose $\underline{H}\in\mathcal{L}$, and for real constants c, $r>0$, $\mu>\mu_o$, and $\delta>0$ the inequality

$$\pm\{|H(\mu+j\omega)-c| - (r\mp\delta)\} \leq 0 \qquad (\omega \text{ real}) \qquad (21)$$

holds, and the μ-shifted Nyquist diagram of H(s) does not encircle the point c; then $\exp^{-\mu}\underline{H}\exp^{\mu}$ is ±conic $(c, r\mp\delta')$ for some constant $\delta'>0$.

Case 2. If $\underline{H}\in\mathcal{L}$ and, for real ω, $\mathrm{Re}\{H(\mu+j\omega)\}\geq a+\delta$ where a, $\mu>\mu_o$ and $\delta>0$ are real constants -- then $\{\exp^{-\mu}\underline{H}\exp^{\mu}-(a+\delta')\underline{I}\}$ is positive for some constant $\delta'>0$.

4.3 A Circle Theorem

The following theorem is the main result of this section; a relationship is obtained between the sector of confinement of N, the logarithmic rate of $\underline{\theta}$, and parameters of disk-shaped permissible regions in the H(s), complex frequency plane (see Fig. 2).

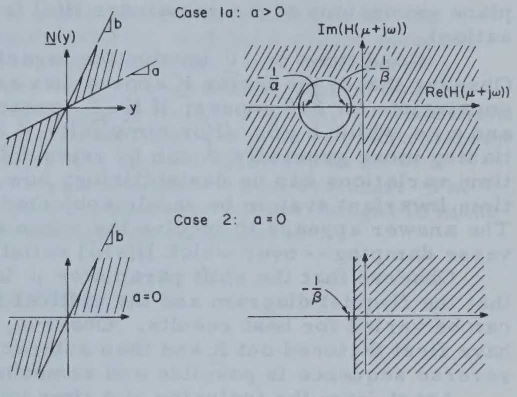


Fig. 2 Admissible Regions for $\underline{N}$ and $\underline{H}$

Theorem 2. Consider equations (4) for the loop of Fig. 1b; the linear operator is factored into two parts, $\underline{H}_1=\underline{K}\,\underline{H}$, and the remaining operator $-\underline{H}_2=\underline{\theta}\,\underline{N}$ consists of a gain $\underline{\theta}$ following nonlinearity $\underline{N}$. Sufficient conditions for $\underline{E}_1$ and $\underline{E}_2$ to be L_∞-bounded are:

(I) $\underline{N}\in\mathcal{N}$, $\theta\in\mathcal{J}$ and

(a) $\underline{\theta}\,\underline{N}$ is confined to sector $\{a+\Delta, b-\Delta\}$ where $b>0$, $a<b$, and $\Delta\geq 0$ are constants.

(b) $\lambda_-\theta\leq\dot{\theta}\leq\lambda_+\theta$ where $\lambda_-<0$ and $\lambda_+>0$ are constants.

(II) $\underline{K}\in\mathcal{RC}$ has no poles more positive than $-p_1$; $\underline{H}\in\mathcal{L}$ satisfies one of the following conditions, for some constant μ where $\mu_o<\mu<0$, and for any real ω:

Case Ia. $a>0$. The μ-shifted Nyquist diagram of H(s) does not encircle $1/2\,(\alpha^{-1}+\beta^{-1})$ and

$$|\underline{H}(\mu+j\omega)+\frac{1}{2}(\alpha^{-1}+\beta^{-1})|\geq\frac{1}{2}|\alpha^{-1}-\beta^{-1}|+\delta \qquad (22)$$

where $\alpha=aK(\infty)$, $\beta=bK(-\lambda)$; $\lambda=\lambda_+[1-(2\mu/\lambda_-)]$ if $2\mu/\lambda_-<1$ and $\lambda=0$ otherwise; δ is a constant greater than zero if $\Delta=0$ and zero otherwise.

Case Ib. $a<0$. The inequality in (22) and the sign of δ are reversed and α becomes $aK(-\lambda)$.

Case 2. $a=0$. $H(\mu+j\omega)\geq-\beta^{-1}+\delta$.

(III) If $K(\infty)=0$ then $a_2=0$ (see Fig. 1b) and $\dot{w}_2\in L_\infty$.

(IV) $(p_1/\lambda_+)+(2\mu/\lambda_-)\geq 1$.

(Note. (Ib) and (IV) can be replaced by the following more general condition: There exists ϕ in $\mathcal{J}$ and $\lambda<p_1$ with the properties that $\phi(t)\exp-\mu t$ is nondecreasing, while $\phi^{-2}(t)\cdot\theta(t)\exp-p_1t$ is nonincreasing.)

Comments. The frequency response of $\underline{H}$ is constrained inside or outside a circle passing some δ away from the points α^{-1}, β^{-1}, -- or if $a = 0$, the circle degenerates to a half-plane. Thus a "critical region" is obtained, which can be viewed as an extension of Nyquist's critical point. The size of this region increases with the size of the nonlinear sector and with the bounds on $\dot{\theta}/\theta$. (Consider, for example, the case $a \geq 0$ with $K(\infty)$ normalized at unity; observe that $K(x)$ is decreasing for $-p_1 < x < 0$; hence α equals a, while β is proportional to b and increasing in λ_+ and $|\lambda_-|$.)

The factor $K(s)$ is typically chosen so as to reduce the left-half-plane excursions of the remainder $H(s)$ (somewhat as in "lead" compensation).

Conditions on $\underline{\theta}$ involve the logarithmic rate $\dot{\theta}/\theta = d/dt(\ell n\ \theta)$. Observe that if the factor $\underline{K}$ approaches an identity (say $p_1 \to \infty$), the constraints on $\dot{\theta}/\theta$ loosen; if $\underline{K} = \underline{I}$ constraints on $\dot{\theta}$ are unnecessary and $a = \alpha$ while $b = \beta$. (For simplicity θ has been postulated differentiable; more generally $\dot{\theta}$ can be replaced by an average rate.) If $\underline{K} \neq \underline{I}$ time variations can be destabilizing; how much varying gain can a time invariant system be safely subjected to without danger of instability? The answer appears to involve the range of negative μ -- i.e., of inverse damping -- over which $H(s+\mu)$ satisfies the circle conditions.

Observe that the shift parameter μ is incompletely specified, and that the Nyquist diagram and the critical region both depend on it; μ can be varied for best results. Observe, also, that where we in effect have first factored out $\underline{K}$ and then subtracted $a\underline{I}$ from $\underline{\theta}\ \underline{N}\underline{K}$ -- the reverse sequence is possible and sometimes more desirable.

Apart from the inclusion of a time varying gain, Theorem 2 differs from Popov's in three main respects: (i) The class of "$\mathcal{RC}$" multipliers is more general than Popov's -- an advantage mainly, it seems, in problems in which a falling open loop frequency response is followed by a rising one (see example); (ii) the nonlinearity carries an additional monotonicity restriction; (iii) constraints on finite order are avoided in the functional setting.

Example. Consider the situation

$$N(x) = \frac{1}{2}\tan^{-1} x; \qquad H_1(s) = \frac{k(s+10)^3}{(s+1)^2(s+100)}$$

$$\theta(t) = 1 + \exp -ct - \exp -dt \qquad (c, d, > 0)$$

It will be shown that the loop is bounded for all $k > 0$. For this purpose let $K(s) = (s+10)/(s+1)$ and obtain a remainder $H(s) = k(s+10)^2/(s+1)(s+100)$. Observe that the following bounds hold

$$\underline{N} \text{ is in } \{0, \tfrac{1}{2}\} \qquad \lambda_- > -c \qquad p_1 = 1$$

$$\underline{\theta} \text{ is in } \{0, 2\} \qquad \lambda_+ < d \qquad \mu_o = -1 + \epsilon$$

Hence $\alpha = 0$ and $\beta = 1 + 9\,(1-d-2\mu dc^{-1})^{-1}$. In order for Theorem 2 to hold it is enough if $c^{-1} + d^{-1} > 1$; if this condition holds then, assuming $\Delta = 0$, $\delta > 0$, boundedness is ensured (Case 2) for

$$\mathrm{Re}\{H(s+\mu)\} \geq \delta - 1/[1+9(1-d-2\mu dc^{-1})^{-1}]$$

An inspection of the pole-zero pattern of $H(s)$ shows that in fact $\mathrm{Re}\{H(s+\mu)\} \geq 0$ for any $-1 < \mu < 0$ and $k \geq 0$.

Observe the role played here by the factor K(s); its removal shifts the remainder H(s) entirely into the right half plane -- without it we could not have obtained $\mathrm{Re}\{H(s+\mu)\} \geq 0$ for all $k>0$; note that the simpler factor $a/(s+a)$, of the Popov type, would not have sufficed -- any real part increase in the range $1<\omega<10$, where $|H_1(s)|$ is falling would have been offset by a decrease in the range $10<\omega<100$ where it is rising. Higher order RC multipliers seem useful mainly when rising and falling ranges alternate.

5. CONCLUSIONS

What emerges here is (1) a fairly inclusive, sufficient but not necessary, theory of input-output stability, and (ii) a circle theorem for loops containing one nonlinearity, multiplier, and time varying gain. The former is a functional theory, resembling that of Nyquist-Bode-Nichols, but differing in concepts from Lyapunov's; its conditions have a qualitative, topological character, and are free of constraints on time invariance, dimensionality, etc.; hence it is likely to be well suited to the analysis of multivariable and distributed systems. The circle theorem, though not necessary, provides enough insight to make possible a reasonably systematic design of stabilizers.

APPENDIX

Proof of Lemma 1

It is enough to show that

$$<x, \underline{\theta}\,\underline{N}\,\underline{K}x;\ t,\sigma> \ \geq\ 0 \qquad (x\in L_{\infty e};\ t\geq 0) \tag{23}$$

for any σ in the range $\sigma > \frac{1}{2}(\lambda - p_1)$. (See proof of Thm. 1 for $\|x; t,\sigma\|$ notation.) It will be shown that the l.h.s. of (23) can be expressed as a sum of components, and that each component is positive; a main ingredient here is the fact that a positive integral on $[0,\infty)$ remains positive after the integrand is weighted with a decreasing function.

For $x\in L_{\infty e}$ let $\underline{K}x = y$; now the theory of partial fraction expansions (for example, in Chap. II of Guillemin's "Synthesis of passive networks") ensures that $K^{-1}(s)$ has an expansion

$$K^{-1}(s) = K^{-1}(-\lambda+2\sigma) + \sum_{i=1}^{n} f_i \frac{s+\lambda-2\sigma}{s+z_i} + f_\infty \cdot (s+\lambda-2\sigma) \tag{24}$$

where $z_i > \lambda - 2\sigma > 0$ for $i=1,2,\ldots n$; $f_i \geq 0$; $f_\infty > 0$ if $K(\infty)=0$ but $f_\infty = 0$ if $K(\infty) > 0$.

Hence x can be expressed as a sum, $x = \sum_{i=0}^{n} \underline{F}_i y$, and it is enough to show that

$$\sum_{i=0}^{n} <\underline{F}_i y,\ \underline{\theta}\,\underline{N}\,y;\ t,\sigma> \ \geq\ 0 \qquad (t\geq 0;\ \sigma > \tfrac{1}{2}(\lambda - p_1)) \tag{25}$$

The components $\underline{F}_i y$ are of one of three types dictated by (24); it will be shown that (25) holds separately for each component type.

Type 1. Here $\underline{F}_o y = \left[K^{-1}(-\lambda+2\sigma)\right] y$. Hence, using the monotonicity of $\underline{N}$ and positivity of θ,

$$<\underline{F}_o y,\ \theta \underline{N} y;\ t,\sigma> \; = \; K^{-1}(-\lambda+2\sigma) <y,\ \theta \underline{N} y;\ t,\sigma> \; \geq \; 0 \qquad (26)$$

Type 2. Here $\underline{F}_\infty y = f_\infty \cdot \left[\dot{y} + (\lambda - 2\sigma) y\right]$ provided $K(\infty) = 0$; hence

$$<\underline{F}_\infty y,\ \theta \underline{N} y;\ t,\sigma> f_\infty^{-1} \; = \; <\dot{y} + (\lambda - 2\sigma) y,\ \theta \underline{N} y;\ t,\sigma>$$

$$= \int_0^t (\lambda - 2\sigma)\ y(\underline{N} y)\ \theta \exp(-2\sigma\tau)\ d\tau + \int_0^t \theta \exp(-2\sigma\tau)\ \dot{y}(\underline{N} y)\ d\tau$$

$$= \int_0^t (\lambda - 2\sigma)\ y(\underline{N} y)\ \theta \exp(-2\sigma\tau)\ d\tau - \int_0^t M(\tau)\ d_\tau(\theta \exp -2\sigma\tau)$$

$$+ \ \theta(t) \exp(-2\sigma t)\ M(t) \qquad (27)$$

where $M(\tau)$ is the result of a partial integration,

$$M(t) \; = \int_0^t \dot{y}\ (\underline{N} y)\ d\tau \; = \int_{y(o)}^{y(t)} N(y)\ dy \qquad (28)$$

Now the last term in (27) is nonnegative because $M(t) \geq 0$ -- because $\underline{N}$ is nondecreasing, and since $y(0) = 0$ -- the last because, for $K(\infty) = 0$, $y(t)$ is the result of an integration on $[0,t]$. Moreover, the difference of the last two integrals in (27) is nonnegative -- since $M \lesssim y(\underline{N} y)$ for $\underline{N}$ monotone -- and since $d_\tau \theta \exp -2\sigma\tau \leq (\lambda - 2\sigma)\ \theta \exp(-2\sigma\tau)\ d\tau$ for $\theta(t) \exp -\lambda t$ nonincreasing.

Type 3. $F_i(s) = f_i \dfrac{s+\lambda-2\sigma}{s+z_i}$. Set $f_i = 1$ and let $u = (\underline{F}_i y - y) \cdot (\lambda - 2\sigma - z_i)^{-1}$; it follows that $\dot{u}$ exists and that $y = \dot{u} + z_i u$ while $\underline{F}_i y = \dot{u} + (\lambda - 2\sigma) u$. Hence

$$<\underline{F}_i y,\ \theta \underline{N} y;\ t,\sigma> \; = \; <\dot{u} + (\lambda - 2\sigma) u,\ \theta \underline{N}(\dot{u} + z_i u);\ t,\sigma> \qquad (29)$$

Subtracting $<\dot{u} + (\lambda - 2\sigma) u,\ \theta \underline{N}(z_i - \lambda + 2\sigma) u;\ t,\sigma>$ from both sides of (29) get

$$<\underline{F}_i y,\ \theta \underline{N} y;\ t,\sigma> \; - \; <\dot{u} + (\lambda - 2\sigma) u,\ \theta \underline{N}(z_i - \lambda + 2\sigma) u;\ t,\sigma>$$

$$= \; <(\dot{u} + z_i u) - (z_i - \lambda + 2\sigma) u,\ \theta \underline{N}(\dot{u} + z_i u) - \theta \underline{N}(z_i - \lambda + 2\sigma) u;\ t,\sigma> \qquad (30)$$

Now observe that the r.h.s. of (30) has the form $<x-y,\ \theta \underline{N} x - \theta \underline{N} y;\ t,\sigma>$ which must be nonegative since $\underline{N}$ is monotone and since $\theta(t) \geq 0$; hence

$$<\underline{F}_i y,\ \theta \underline{N} y;\ t,\sigma> \; \geq \; <\dot{u} + (\lambda - 2\sigma) u,\ \theta \underline{N}(z_i - \lambda + 2\sigma) u;\ t,\sigma> \; \geq \; 0 \qquad (31)$$

The positivity of (31) is proved by the argument used in (27), and using the fact that $z_i - \lambda + 2\sigma > 0$. QED.

Proof of Lemmas 1a and 1b

First suppose $a = 0$. Apply Lemma 1 to the operator $\phi^{-2} \theta \underline{N} \underline{K}$; using (26) and the fact that $\theta \underline{N}$ is in $\{a,\ b-\Delta'\}$ conclude that

$$<\phi^{-2}\underline{\theta}\,\underline{N}\,\underline{K}x,\ \underline{\theta}\,\underline{N}\,\underline{K}x-(b-\Delta')K(-\lambda+2\sigma)x;\ t,\sigma> \ \leq 0 \quad (x\in L_{\infty e};\ t\geq 0) \qquad (32)$$

for $\sigma > \frac{1}{2}(\lambda - p_1)$. For any $y \in L_{\infty e}$, let $x = \underline{\phi}y$; hence $x \in L_{\infty e}$, so (32) gives

$$<\underline{\phi}^{-1}\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi}y,\ \underline{\phi}^{-1}\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi}y-(b-\Delta')K(-\lambda+2\sigma)y;\ t,\sigma> \ \leq 0 \quad (t\geq 0) \qquad (33)$$

Now $K(-\lambda+2\sigma)$ is continuous in σ near the origin; hence for σ near enough to zero there exists $\Delta>0$ for which $(b-\Delta')K(-\lambda+2\sigma) = bK(-\lambda)-\Delta$. Therefore (33) implies that $\underline{\phi}^{-1}\,\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi}$ is in $\{0, bK(-\lambda)-\Delta\}$.

Next suppose $a>0$. It can be shown that $\underline{N}$ is in $\{a_N, b_N\}$ while θ is in $\{a_\theta, b_\theta\}$ where $a_\theta a_N = a$ and $b_\theta b_N = b$. Decompose $\underline{\phi}^{-1}\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi}$ into three parts,

$$\underline{\phi}^{-1}\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi} = \underline{\phi}^{-1}\underline{\theta}(\underline{N}-a_N\underline{I})\underline{K}\,\underline{\phi} + a_N\underline{\phi}^{-1}\underline{\theta}(\underline{K}-K(\infty)\underline{I})\underline{\phi} + a_N K(\infty)\underline{\phi}^{-1}\underline{\theta}\,\underline{\phi} \qquad (34)$$

Now the first two terms on the r.h.s. of (34) can be confined using the $a=0$ rule, and are in $\{0,\ b_\theta(b_N-a_N)K(-\lambda) - \Delta_1\}$ and in $\{0,\ a_N b_\theta[K(-\lambda)-K(\infty)] - \Delta_2\}$ respectively. The last term on the r.h.s. of (34) is in $\{a_N a_\theta K(\infty),\ a_N b_\theta K(\infty)\}$. Lemma 1a now follows by the following sum rule for sectors: "If $\underline{H}$ and $\underline{K}$ are in $\{a,b\}$ and $\{c,d\}$ respectively, $(\underline{H}+\underline{K})$ is in $\{a+c,\ b+d\}$."

Finally if $a<0$, decompose $\underline{N}$ into a difference $\underline{N} = \underline{N}_+ - \underline{N}_-$ of two nondecreasing parts, such that $\underline{\theta}\,\underline{N}_+$ is in $\{0,b\}$ and $\underline{\theta}\,\underline{N}_-$ in $\{0,a\}$; Lemma 1b follows by the sum rule.

Proofs of Lemma 2 and Theorem 2

Limitations of space prevent us from completing these proofs.

Lemma 2 is available in essence in a report entitled "On the Input-Output Stability of Time Varying Nonlinear Feedback Systems -- II"; the proof involves Parseval's Theorem, Theorem 5 of Paley-Wiener's "Fourier Transforms in the Complex Domain", and the principle of the argument.

Theorem 2 is an application of Theorem 1. First a factor $\underline{\phi}$ is formed; $\phi = \theta^{\mu/\lambda_-}$ if $\mu/\lambda_- \leq 1$ or $\phi = \theta^{1/2}$ otherwise; it can be shown that ϕ satisfies the conditions mentioned in Sect. 4.3. Next the loop equations are transformed; the new loop has the form of Fig. 1 with $\underline{H}_1 = \underline{\phi}^{-1}\underline{H}\,\underline{\phi}$ and $\underline{H}_2 = -\underline{\phi}^{-1}\underline{\theta}\,\underline{N}\,\underline{K}\,\underline{\phi}$; $w_1 + a_1x$ is multiplied by $\underline{\phi}^{-1}$; $w_2 + a_2x$ is moved ahead of $\underline{K}\,\underline{\phi}$ and replaced by $w_2' + a_2'x$ where $\underline{K}\,\underline{\phi}(w_2' + a_2'x) = w_2 + a_2x$. Finally Lemma 2 shows that $\underline{H}_1$ is in $\{\alpha, \beta\}$, while Lemma 1 shows that $\underline{H}_2$ is outside $\{-\alpha^{-1}, -\beta^{-1}\}$, etc.

REFERENCES

[1] Popov, V. M., "Absolute Stability of Nonlinear Systems of Automatic Control," Automation and Remote Control, March 1962, Vol. 22, No. 8, pp. 857-875 (Russian original in August 1961).

[2] Narendra, K. S., and Goldwyn, R. M., "A Geometrical Criterion for the Stability of Certain Nonlinear Nonautonomous Systems," IEEE Trans. on Circuit Theory, September 1964, pp. 406-408.

[3] Kudrewicz, J., "Stability of Nonlinear Feedback Systems," Automat. i Telemech., 1964, Vol. XXV, No. 8.

[4] Zames, G., (a) "On the Stability of Nonlinear, Time Varying Feedback Systems," Proceedings of the National Electronics Conference, October, 1964, Vol. 20, pp. 725-730. (b) "Conditions for the Input-Output Stability of Nonlinear, Time Varying Systems, Derived Using the Concept of Loop Gain," ESL Report, March 1, 1965.

[5] Brockett, R. W. (a) and Willems, J. W., "Frequency Domain Stability Criteria, Parts I and II," Proceedings of the JACC, 1965, pp. 735-737. (b) and Forys, L. J., "On the Stability of Systems Containing a Time Varying Gain," Proc. 1964 Allerton Conference.

RECURSIVE ESTIMATION OF THE MODAL TRAJECTORY

ROBERT E. LARSON and JOHN PESCHON
Stanford Research Institute, Menlo Park, California

ABSTRACT

A recursive equation based on dynamic programming for finding the most likely trajectory of a dynamic process observed through a noisy measurement system is presented; the procedure can be applied to non-linear processes with non-Gaussian noise. It differs from the recently developed Bayesian estimation procedure in that the most likely estimate of the entire trajectory up to the present time, rather than of the present state only, is generated. It is shown that the two procedures in general yield different estimates of the present state; however, in the case of linear systems with Gaussian noise, both procedures reduce to the Kalman-Bucy filter. Illustrative examples are presented, and the modal trajectory procedure is compared with the Bayesian procedure and with other estimation techniques in terms of computational requirements and applicability.

I MODAL TRAJECTORY ESTIMATION PROCEDURE

A. Problem Formulation

The problem of optimum estimation is formulated, with reference to Fig. 1, as follows:

Given:
- The known difference equation describing the signal-generating process* $x_{k+1} = f\ (x_k, w_k, k)$. (1)
- The initial probability density $p\ (x_0)$.
- The statistics of w, $p\ (w_i)$, $i = 0, 1, \ldots, k$.
- The known relation describing the measurement system*

$$z_k = k\ (x_k, v_k). \qquad (2)$$

- The statistics of v, $p\ (v_i)$, $i = 1, \ldots, k$.
- The measurements z_i, $i = 1, \ldots, k$.

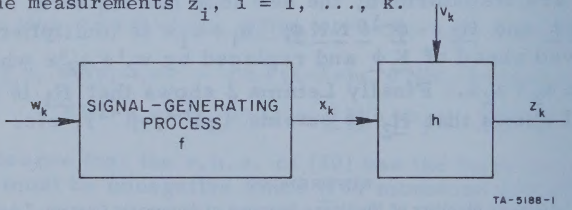


FIG. 1 OPTIMUM ESTIMATION

x_k: state of signal generating process of kth sampling time
w_k: random perturbation at time k
v_k: measurement noise at time k
z_k: measurement at time k.

---

*In general, the variables x_k, u_k, w_k, v_k, and z_k are vectors. The dimension of x_k usually exceeds that of z_k; the dimension of v_k is assumed equal to that of z_k, i.e., each measurement is affected by noise.

Find: The maximum likelihood estimate of the sequence of states $x_0, x_1, \ldots, x_k$.

The sequence of states is called the trajectory, and is abbreviated as X_k.

$$X_k = \{x_0, x_1, \ldots, x_k\}. \tag{3}$$

The maximum likelihood estimate is written as

$$\hat{X}_{k/k} = \{\hat{x}_{0/k}, \hat{x}_{1/k}, \ldots, \hat{x}_{k/k}\}. \tag{4}$$

The sequence of observations is abbreviated as

$$Z_k = \{z_1, z_2, \ldots, z_k\}. \tag{5}$$

The modal trajectory is that trajectory which maximizes the conditional probability distribution $p\,(X_k|Z_k)$, where

$$p\,(X_k|Z_k) \;=\; p\,(x_0, x_1, \ldots, x_k|z_1, z_2, \ldots, x_k). \tag{6}$$

The probability of the modal trajectory estimate can thus be written as

$$p\,(\hat{X}_{k/k}|Z_k) = \underset{X_k}{\text{Max}}\, \{p(X_k|Z_k)\}. \tag{7}$$

One approach to finding this estimate is to compute $p(X_k|Z_k)$ for all possible X_k and then to search over all of these probabilities until the maximum value is found. The estimate $\hat{X}_{k/k}$ is the particular sequence X_k for which the maximum is attained.

B. Recursive Equations

The ideas of dynamic programming[1*] can be brought to bear on this problem in such a way that the simultaneous minimization over $X_k = \{x_0, x_1, \ldots, x_k\}$ in Eq. (7) is reduced to a sequential minimization. First, Eq. (7) is rewritten as

$$p\,(\hat{X}_{k/k}|Z_k) = \left[\underset{x_k}{\text{Max}} \quad \underset{X_{k-1}}{\text{Max}} \;\{p(X_k|Z_k)\}\right]. \tag{8}$$

The result of the maximization over X_{k-1} is the probability of the maximum likelihood trajectory which terminates at a specified value of x_k. This probability is denoted as $I(x_k, k)$

$$I(x_k, k) \;=\; \underset{X_{k-1}}{\text{Max}}\, \{p(X_k|Z_k)\}. \tag{9}$$

The desired recursive relation is obtained by using Bayes' theorem[2] to express $I(x_{k+1}, k+1)$ in terms of $I(x_k, k)$. From Bayes' theorem it follows that

$$p(X_{k+1}/Z_{k+1}) = \frac{p(z_{k+1}|X_{k+1}, Z_k)\; p(X_{k+1}|Z_k)}{p(z_{k+1}|Z_k)}. \tag{10}$$

Since the observation z_{k+1} depends only on x_{k+1}, Eq. (10) is simplified to

$$p(X_{k+1}|Z_{k+1}) = \frac{p(z_{k+1}|x_{k+1})\; p(X_{k+1}|Z_k)}{p(z_{k+1}|Z_k)}. \tag{11}$$

---

[*] References are listed at the end of the paper.

The term $p(X_{k+1}|Z_k)$ is next broken down into

$$p(X_{k+1}|Z_k) = p(x_{k+1}|x_k)\ p(X_k|Z_k), \tag{12}$$

with the result that $I(x_{k+1},\ k+1)$ can be written as

$$\begin{aligned} I(x_{k+1},\ k+1) &= \max_{X_k} \left\{ \frac{p(z_{k+1}|x_{k+1})\ p(x_{k+1}|x_k)}{p(z_{k+1}|Z_k)}\ p(X_k|Z_k) \right\} \\ &= \max_{x_k} \left\{ \frac{p(z_{k+1}|x_{k+1})\ p(x_{k+1}|x_k)}{p(z_{k+1}|Z_k)}\ I(x_k,\ k) \right\} \end{aligned} \tag{13}$$

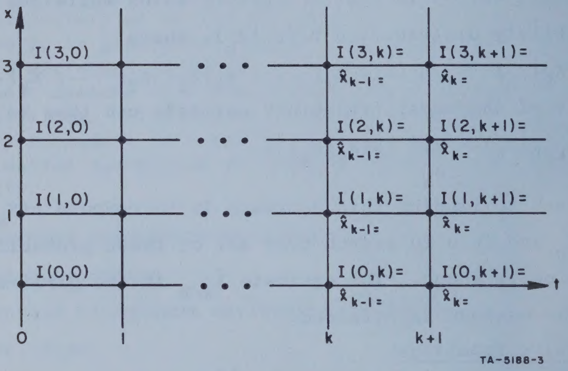


FIG. 2 COMPUTATIONAL PROCEDURE TO IMPLEMENT (13) OR (14)

Equation (13) constitutes the desired recursive relation. The probability density function $p(z_{k+1}|x_{k+1})$ is obtained from Eq. (2) and knowledge of $p(v_{k+1})$, while $p(x_{k+1}|x_k)$ is found from Eq. (1) and knowledge of $p(x_k)$. The function $p(z_{k+1}|Z_k)$ is used for normalization purposes only and need not be carried along; thus, Eq. (13) can be rewritten as

$$I^*(x_{k+1},\ k+1) = \max_{x_k} \left\{ p(z_{k+1}|x_{k+1})\ p(x_{k+1}|x_k)\ I^*(x_k,\ k) \right\}, \tag{14}$$

where the function I^* is proportional, but not equal to the function I. Because of this proportionality, the same value of x_k minimizes both Eqs. (13) and (14). This value will be denoted as $\hat{x}_k(x_{k+1},\ k+1)$.

The modal trajectory $X_{k/k}$ is determined by first using Eq. (8) in the form

$$p(\hat{X}_{k/k}|Z_k) = \max_{x_k} \left\{ I(x_k,\ k) \right\}. \tag{15}$$

The value of x_k for which the maximum is attained in Eq. (15) is $\hat{x}_{k/k}$, the last state along the modal trajectory. The same value of $\hat{x}_{k/k}$ is determined if $I^*(x_k,\ k)$ is maximized rather than $I(x_k,\ k)$ in Eq. (15). The rest of the trajectory is then found by tracing backward a stage at a time, using the function $\hat{x}_k(x_{k+1},\ k+1)$. Formally,

$$\hat{x}_{i/k} = \hat{x}_i(\hat{x}_{k+1/k}, i+1), \tag{16}$$

where $\hat{x}_{k/k}$ is used to initiate this procedure.

A computational procedure completely analogous to forward dynamic programming[3] can be used to compute the functions $I(x_k, k)$ [or equivalently $I^*(x_k, k)$] and $\hat{x}_{k-1}(x_k, k)$ from Eq. (14) [or Eq. (15)]. This procedure is illustrated in Fig. 2, and the major steps are listed below.

(1) Quantize the space $[x, t]$ to obtain a grid consistent with the accuracy requirements of the problem.

(2) Initialize by defining $I(x_0, 0) = I^*(x_0, 0) = p(x_0)$, the a priori distribution.

(3) For each quantized state x_1, calculate $I(x_1, 1)$ [or $I^*(x_1, 1)$] from z_1 and Eq. (13) [or Eq. (14)].

(4) Write $\hat{x}_0(x_1, 1)$ as the value of x_0 for which Eq. (13) [or Eq. (14)] is maximized in the above calculation.

(5) Repeat Steps (3) and (4) for each stage until stage k is reached.

(6) Determine the modal trajectory $\hat{X}_{k/k}$ by first using Eq. (15) to find $\hat{x}_{k/k}$, then iteratively applying Eq. (16) to the function $\hat{x}_i(x_{i+1}, i+1)$.

C. Derivation of the Modal Trajectory for the Linear Gaussian White Noise Case

In the linear Gaussian white noise case, Eqs. (1) and (2) are rewritten as

$$x_{k+1} = \Phi_k x_k + \Gamma_k w_k \tag{17}$$

$$z_k = H_k x_k + v_k. \tag{18}$$

The probability density functions $p(x_0)$, $p(w_k)$ and $p(v_k)$ are Gaussian with means $\hat{x}_0$, 0 and 0 and with covariances P_0, Q_k and R_k, respectively. It is assumed that

$$\begin{aligned} E[v_i\, w_j^T] &= 0 , && \text{all i and j} \\ E[v_i\, v_j^T] &= 0 , && i \neq j \\ E[w_i\, w_j^T] &= 0 , && i \neq j \\ E[x_0\, w_j^T] &= E[x_0, v_i{}^T] = 0 && \text{All i and j} \end{aligned} \tag{19}$$

It is well known[4] that, in this case, the probability density functions $p(z_{k+1}|x_{k+1})$, $p(x_{k+1}|x_k)$ and $p(X_k|Z_k)$ are all Gaussian. Using the white noise assumption, Eq. (19), $I(x_k, k)$ can be written as

$$\begin{aligned} I(x_k, k) &= \underset{x_0, \ldots, x_{k-1}}{\text{Max}} \left\{ c_1 \exp -\frac{1}{2} \left[\sum_{i=0}^{k} (x_k - \bar{x}_{k/k})^T P_{k/k}^{-1} (x_k - \bar{x}_{k/k}) \right] \right\} \\ &= c_1 \exp -\frac{1}{2} [(x_k - \bar{x}_{k/k})^T P_{k/k}^{-1} (x_k - \bar{x}_{k/k}) + c_2], \end{aligned} \tag{20}$$

where $\bar{x}_{i/k}$ is the conditional mean, $P_{i/k}$ is the conditional covariance, c_1 is a normalization constant, and c_2 is a constant independent of x_k.

The probability density functions $p(z_{k+1}|x_{k+1})$ and $p(x_{k+1}|x_k)$ are characterized by*

$$\begin{aligned} E\,(z_{k+1}|x_{k+1}) &= H\,x_{k+1} \\ \mathrm{Cov}\,(z_{k+1}|x_{k+1}) &= R \end{aligned} \qquad (21)$$

$$\begin{aligned} E\,(x_{k+1}|x_k) &= \Phi\,x_k \\ \mathrm{Cov}\,(x_{k+1}|x_k) &= \Gamma\,Q\,\Gamma^T. \end{aligned} \qquad (22)$$

The recursive relation (13) then becomes

$$\begin{aligned} I(x_{k+1},k+1) = \underset{x_k}{\mathrm{Max}}\Big\{c_1 \exp\ -\tfrac{1}{2}\Big[&(z_{k+1} - H\,x_{k+1})^T R^{-1}(z_{k+1} - H\,x_{k+1}) \\ &+ (x_{k+1} - \Phi x_k)^T(\Gamma Q \Gamma^T)^{-1}(x_{k+1} - \Phi x_k) \\ &+ (x_k - \bar{x}_{k/k})^T P^{-1}_{k/k}(x_k - \bar{x}_{k/k}) + c_2\Big]\Big\}. \end{aligned} \qquad (23)$$

Maximization of an exponential is equivalent to minimization of the negative exponent. The latter minimization can be performed by differentiation with respect to x_k or by completion of the square. The result is

$$I(x_{k+1},k+1) = c_1 \exp\ -\tfrac{1}{2}\ [(x_{k+1} - \bar{x}_{k+1/k+1})^T P^{-1}_{k+1/k+1}(x_{k+1} - \bar{x}_{k+1/k+1}) + c_2], \qquad (24)$$

where

$$\begin{aligned} \bar{x}_{k+1/k+1} &= \Phi\,\bar{x}_{k/k} + P_{k+1/k+1}\,H^T R^{-1}(z_{k+1} - H\Phi\bar{x}_{k/k}) \\ (P^{-1}_{k+1/k+1} &= H^T R^{-1} H + (\Phi\,P_{k/k}\,\Phi^T + \Gamma\,Q\,\Gamma^T)^{-1}. \end{aligned} \qquad (25)$$

From Eq. (24) it is clear that

$$\hat{x}_{k+1/k+1} = \bar{x}_{k+1/k+1}. \qquad (26)$$

Since Eqs. (25) are the standard equations for the Kalman-Bucy filter,[4] it follows that this estimate of the last state along the trajectory is the same as the Kalman-Bucy estimate. It can also be shown that the rest of the trajectory estimate is the same as the Kalman-Bucy smoothing solution.[4,5]

II THE BAYESIAN ESTIMATION PROCEDURE

A. Problem Formulation

A different procedure, usually called the Bayesian estimation procedure,[4-6] has been developed for finding the maximum likelihood estimate of the present state. The given data are exactly the same as in Sec. I-A, but the objective is

Find: The maximum likelihood estimate of the present state x_k.

This estimate will be written as $\hat{\hat{x}}_{k/k}$ to distinguish it from $\hat{x}_{k/k}$, the estimate of the last state along the trajectory $\hat{\hat{X}}_{k/k}$.

---

*For simplicity of notation, the argument of the functions H, Φ, Γ, R, and Q will be left out, but it is understood that they are all allowed to vary with time.

B. Recursive Equation

The recursive relation is obtained by applying Bayes' theorem in the form

$$p(x_{k+1}|z_{k+1}) = \frac{p(z_{k+1}|x_{k+1}, Z_k)\ p(x_{k+1}|Z_k)}{p(z_{k+1}|Z_k)} \ . \tag{27}$$

Using Eq. (11) and noting that

$$p(x_{k+1}|Z_k) = \int_{x_k} p(x_{k+1}|x_k)\ p(x_k|Z_k)\ d\ x_k, \tag{28}$$

where $\int_{x_k} \cdots d\ x_k$ denotes integration over all possible values of x_k, Eq. (27) becomes

$$p(x_{k+1}|Z_{k+1}) = \int_{x_k} \frac{p(z_{k+1}|x_{k+1})\ p(x_{k+1}|x_k)}{p(z_{k+1}|Z_k)}\ p(x_k/Z_k)\ d\ x_k . \tag{29}$$

This equation determines the probability density function of the state x_{k+1} conditioned on the observation sequence Z_{k+1} in terms of the known probability densities $p(v_{k+1})$ and $p(w_k)$, and the just computed probability density function of x_k conditioned on Z_k. The term $p(z_{k+1}|Z_k)$ is just a normalization factor. The a priori distribution $p(x_0)$ is used as an initial condition for Eq. (29). The maximum likelihood estimate is obtained from

$$p(\hat{\hat{x}}_{k/k}) = \operatorname*{Max}_{x_k}\ [p(x_k|Z_k)], \tag{30}$$

where $\hat{\hat{x}}_{k/k}$ is the value of x_k that maximizes Eq. (30).

In the linear Gaussian white noise case, it can be shown[4-6] that the Kalman-Bucy equations are again obtained. Thus, in this particular case the estimates $\hat{\hat{x}}_{k/k}$ and $\hat{x}_{k/k}$ are identical. However, it will be shown in the next section that this is not true in general.

Although this procedure was developed primarily for estimating x_k, it can be extended to find the estimate of prior states. These estimates are commonly called the smoothing solution[4,5,7], and the set of all of these estimates will be denoted as

$$\hat{\hat{X}}_{k/k} = \{\hat{\hat{x}}_{0/k}, \hat{\hat{x}}_{1/k}, \ldots, \hat{\hat{x}}_{k/k}\}. \tag{31}$$

Unfortunately, considerably more computations are required to obtain this solution over and above those used in obtaining $\hat{\hat{x}}_{k/k}$. The most efficient approach is to first compute $\hat{\hat{x}}_{i/i}$, $i = 1,2,\ldots,k$, and then to apply iteratively the equation

$$p(x_{i-1}, x_i|Z_k) = p(x_i|x_{i-1})\ p(x_{i-1}|Z_{i-1})\ p(z_i \ldots z_k|x_i). \tag{32}$$

The probability density function $p(x_{i-1}|Z_{i-1})$ is obtained during the computation of $\hat{\hat{x}}_{k/k}$ using Eq. (29). The function $p(x_i|x_{i-1})$ uses Eq. (1) and $p(w_{i-1})$, while $p(z_i, \ldots z_k|x_i)$ uses both of Eqs. (1) and (2) and $p(w_j)$ and $p(v_j)$, $j = i, i+1, \ldots k$.

The estimate $\hat{\hat{x}}_{i-1/k}$ is then obtained from

$$p(\hat{\hat{x}}_{i-1/k}) = \underset{x_{i-1}}{\mathrm{Max}} \left[\int_{x_i} p(x_{i-1}, x_i | Z_k)\, d\, x_i \right] \tag{33}$$

Explicit equations can be obtained for the linear Gaussian white noise case[5,6]. These equations are again the Kalman-Bucy smoothing solution, just as in the modal trajectory calculation.

III ILLUSTRATIVE EXAMPLES USING BOTH PROCEDURES

A. Counter-Example to Show that $\hat{\hat{x}}_{k/k} \neq \hat{x}_{k/k}$

Although the estimates are the same for both procedures in the linear Gaussian white noise case, this is not true in general. As a simple counter-example, consider the two-stage process shown in Fig. 3. The state of the system is assumed to be one of two values, 0 and 1. The a priori probabilities of the two states are $p(x_0 = 0) = 0.4$ and $p(x_0 = 1) = 0.6$.

The system equations are such that $p(x_1 | x_0)$ is given in Table I. The observation at k=1 is also assumed to have two values, 0 and 1. The observation equations specify $p(z_1 | x_1)$ as in Table II.

$p(x_1 | x_0)$

$x_1 \backslash x_0$	0	1
0	0.9	0.3
1	0.1	0.7

Table I System Equation

$p(z_1 | x_1)$

$z_1 \backslash x_1$	0	1
0	0.5	0.5
1	0.5	0.5

Table II Observation Equation

If the observation $z_1 = 1$ is received, the conditional probabilities of the four possible trajectories are as shown in Fig. 3 (the same results are achieved if $z_1 = 0$).

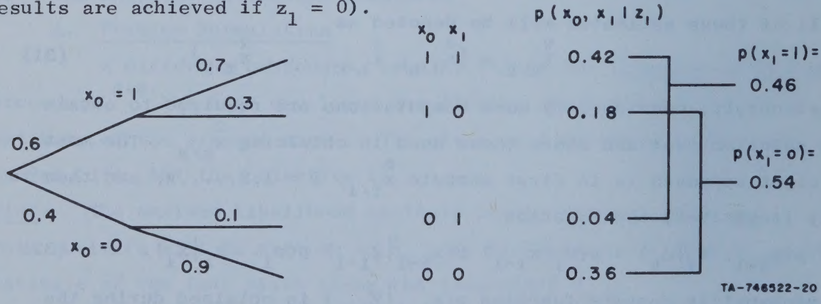


FIG. 3 COUNTEREXAMPLE TO SHOW THAT $\hat{x}_{k/k} \neq \hat{\hat{x}}_{k/k}$

The maximum likelihood trajectory is found to be $x_0 = 1$, $x_1 = 1$, which has probability 0.42. On the other hand, the probability that $x_1 = 1$, which is obtained by adding the probabilities of trajectories $x_0 = 1$, $x_1 = 1$, and $x_0 = 0$, $x_1 = 1$, is 0.46, while the probability that $x_1 = 0$ is 0.54. Thus, the maximum likelihood value of the present state is $x_1 = 0$. In this simple example, the final state along the maximum likelihood trajectory hence differs from the maximum likelihood estimate of this state, i.e., in this case, $\hat{x}_{1/1} \neq \hat{\hat{x}}_{1/1}$.

This example also clarifies the relation between these two estimates. The maximum likelihood trajectory is the one with the largest probability, while the maximum likelihood final state is found by adding the probabilities of all trajectories with the same final state and choosing the value with the largest probability.

B. A Second Nonlinear Non-Gaussian Example

The state again takes on only the two values, 0 and 1. The system equation is determined by Table III where $p(w_k = 0) = 0.875$ and $p(w_k = 1) = 0.125$. The observation equation is specified by Table IV, where $p(v_k = 0) = 0.75$ and $p(v_k = 1) = 0.25$.

x_{k+1}

w_k \ x_k	0	1
0	0	1
1	1	0

Table III Truth Table for System Equation

z_k

v_k \ x_k	0	1
0	0	1
1	1	1

Table IV Truth Table for Observation Equation

The a priori distribution is $p(x_0 = 0) = 0.2$, $p(x_0 = 1) = 0.8$.

In Table I the maximum likelihood trajectory and the maximum likelihood state are both computed. Values of w_k and v_k were generated using a random number generator. The functions $I^*(0,k)$ and $I^*(1,k)$ were normalized so that $I^*(0,k) = 1$ always. The maximum likelihood trajectory after 25 measurements, $\hat{X}_{25/25}$, is shown along with the sequence of estimates $\hat{\hat{x}}_{k/k}$, $k = 1,2,\ldots,25$. The maximum likelihood trajectory is in error only at one point, while the sequence $\hat{\hat{x}}_{k/k}$ is in error four times.

IV CONCLUSIONS

In a given problem with fixed quantization, if an estimate of only the present state is desired, then the modal trajectory estimation procedure requires almost exactly the same amount of computations as the Bayesian procedure. Experimentation has shown that these estimates are about equally accurate; in most cases they are the same.

Table V

MAXIMUM LIKELIHOOD TRAJECTORY AND MAXIMUM LIKELIHOOD ESTIMATE
OF THE SEQUENCE OF STATES IN ONE-DIMENSIONAL EXAMPLE

(Errors in the Estimates are Circled.)

					Maximum Likelihood Trajectory Calculation					Bayesian Calculation		
k	x(k)	w(k)	v(k)	z(k)	$I^*(0,k)$	$\hat{x}_{k-1}(0,k)$	$I^*(1,k)$	$\hat{x}_{k-1}(1,k)$	$\hat{x}_{k/25}$	p[x(k)=0]	p[x(k)=1]	$\hat{x}_{k/k}$
0	1	0			1	-	4	-	1	0.200	0.800	1
1	1	1	1	1	1	0	16	1	1	0.087	0.913	1
2	0	0	1	1	1	1	28	1	①	0.055	0.945	①
3	0	0	0	0	1	1	0	-	0	1.000	0.000	0
4	0	0	1	1	1	0	0.57	0	0	0.636	0.364	0
5	0	0	1	1	1	0	2.29	1	0	0.274	0.726	①
6	0	0	0	0	1	0	0	-	0	1.000	0.000	0
7	0	0	0	0	1	0	0	-	0	1.000	0.000	0
8	0	1	0	0	1	0	0	-	0	1.000	0.000	0
9	1	0	0	1	1	0	0.57	0	1	0.636	0.364	⓪
10	1	0	0	1	1	0	2.29	1	1	0.274	0.726	1
11	1	0	0	1	1	0	9.14	1	1	0.111	0.889	1
12	1	0	1	1	1	1	28	1	1	0.062	0.938	1
13	1	0	0	1	1	1	28	1	1	0.048	0.952	1
14	1	0	1	1	1	1	28	1	1	0.046	0.954	1
15	1	0	0	1	1	1	28	1	1	0.045	0.955	1
16	1	1	0	1	1	1	28	1	1	0.045	0.955	1
17	0	0	0	0	1	1	0	-	0	1.000	0.000	0
18	0	0	0	0	1	0	0	-	0	1.000	0.000	0
19	0	0	0	0	1	0	0	-	0	1.000	0.000	0
20	0	0	1	1	1	0	0.57	0	0	0.636	0.364	0
21	0	1	0	0	1	0	0	-	0	1.000	0.000	0
22	1	0	0	1	1	0	0.57	0	1	0.636	0.364	⓪
23	1	0	0	1	1	0	2.29	1	1	0.274	0.726	1
24	1	0	0	1	1	0	9.14	1	1	0.111	0.889	1
25	1		1	1	1	1	28	1	1	0.062	0.938	1

If an entire sequence of states is desired, it is generally much easier to recover it from the modal trajectory calculations than to use the Bayesian smoothing solution. In the former case, the trajectory is found by examining the results of the calculations already performed, while in general the latter case requires computation of a whole new sequence of probability distributions. Experimentation has shown that the two smoothed estimates are about equally accurate, while the sequence of unsmoothed Bayesian estimates $\hat{\hat{x}}_{i/i}$ is much less accurate than the corresponding points on the modal trajectory $\hat{X}_{k/k}$ (see Table V).

Both procedures can be applied to the nonlinear, non-Gaussian, non-white noise case. In the linear, Gaussian white noise case, both procedures reduce to the Kalman-Bucy filter and smoother.

Both procedures are usually more accurate than polynomial fitting procedures[5] because they can use models of the process being estimated. However, they generally require more computational effort to implement.

In summary, as a technique for estimating the state, the modal trajectory procedure is comparable to the Bayesian procedure in general applicability, accuracy, and computational requirements. It generally is more efficient than the Bayesian smoothing procedure in obtaining "after the fact" estimates of the entire trajectory.

REFERENCES

1. Bellman, R.E., and Dreyfus, S.E., Applied Dynamic Programming, (Princeton University Press, Princeton, New Jersey, 1962).
2. Parzen, E., Modern Probability Theory and Its Applications, (John Wiley and Sons, Inc., New York, 1960).
3. Larson, R.E., State Increment Dynamic Programming, Chapter 9 (Academic Press, New York, New York, 1965).
4. Kalman, R.E. and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Trans. A.S.M.E., J. Basic Engineering, Vol. 83 No. 1, pp. 95-108 (March 1961).
5. Lee, R.C.K., Optimal Estimation Identification and Control (MIT Press, Cambridge, Massachusetts, 1964).
6. Peschon, John, "Review and Extension of Real-Time Estimation Techniques," submitted to the Third IFAC Congress, London, 1966.
7. Rauch, H.E., Tung, F., and Striebel, C.T., "Maximum Likelihood Estimates of Linear Dynamic Systems," AIAA Journal, Vol. 3, No. 8, pp. 1445-1450 (August 1965).

NONLINEAR PROGRAMMING FOR OPTIMUM CONTROL

JOHN PESCHON
Stanford Research Institute, Menlo Park, California

ABSTRACT

Many constrained static and dynamic optimization problems reduce to a nonlinear programming procedure, to which the Kuhn and Tucker theorem provides an efficient algorithm for numerical solution.

The static problem of optimally allocating active and reactive powers subject to voltage, angle, power and reactive power constraints in an electric power system is considered. The ability of the resulting algorithm to handle high-order systems is discussed and the importance of the "dual variables" for planning and tarification purposes is stressed.

The procedure is next modified to encompass dynamic optimization problems; as an illustration, the optimum long-term planning of an electric power system is discussed.

I POWER SYSTEM OPTIMIZATION

An electric power system of N + 1 nodes is considered. There may be a production P_i, Q_i and a consumption C_i, D_i of active and reactive power respectively, at the ith node, which is further characterized by its voltage U_i and phase angle θ_i. For pure production nodes $C_i = D_i = 0$, and for pure consumption nodes $P_i = Q_i = 0$. A frequent intermediate case is $P_i = 0$ and $Q_i \neq 0$.

The total productions $\sum_i P_i$ and $\sum_i Q_i$ must equal the total consumptions $\sum_i C_i$ and $\sum_i D_i$ plus the respective transmission losses. The cost J, in $ per unit time, of operating the system depends only on the active powers P_i, the reactive powers Q_i being free once the equipment required for their production (e.g., capacitors) has been installed. Thus

$$J = F(P_0,\ldots,P_N). \tag{1}$$

If, in a given power system, the P_i, C_i, Q_i and D_i are specified, then the voltages U_i and the phase angles θ_i between these voltages and U_0 (whose phase angle is arbitrarily set equal to zero) are fixed.

The related variables P_i, Q_i, and U_i are subject to upper and lower constraints, e.g.,

$$P_i^m \leq P_i \leq P_i^M. \tag{2}$$

The phase differences $\theta_i - \theta_\alpha (\alpha = 0,1,\ldots,i-1,i+1,\ldots,N)$ between adjacent nodes are similarly constrained.

The problem of power system optimization then consists of minimizing J subject to three sets of constraints, namely

(1) Satisfaction of demand C_i and D_i;

(2) satisfaction of the constraints imposed upon P_i, Q_i, U_i and $\theta_i - \theta_\alpha$;

(3) satisfaction of the network relations which must exist between the variables P_i, Q_i and U_i, θ_i.

The stated problem is one of static optimization in the presence of constraints. The optimization conditions provided by the theorem of Kuhn and Tucker constitute a practical solution, which can be implemented for real-time control or used for planning purposes.

The above formulation of the constrained power system optimization problem and its solution by the Kuhn and Tucker theorem is due to Carpentier[1*] and Carpentier and Siroux[2]. The remainder of Part I is largely a condensation of their work, which is not available in English translation. A program for the numerical solution of a variable voltage problem has been established at Stanford Research Institute and is described in Sec. H. The shortcomings of previously known power system optimization procedures, notably Lagrange's method and the method of differential losses, are discussed in Ref. 1 and will not be repeated here.

A. The Theorem of Kuhn and Tucker[3]

Given: (1) A cost $F(\underline{x})$ of the vector variable $\underline{x}$

(2) Inequality constraints of the form

$$\underline{G}(\underline{x}) \le 0 \tag{3}$$

(3) Equality constraints

$$\underline{H}(\underline{x}) = 0 \tag{4}$$

and assuming convexity for F, G and H, the conditions under which $F(\underline{x})$ is minimum is d L = 0 with

$$L = F(\underline{x}) + \underline{\alpha}^T \underline{G}(\underline{x}) + \underline{\beta}^T \underline{H}(\underline{x}) \tag{5}$$

$$\alpha_i \ge 0, \ \beta_i \text{ arbitrary} \tag{6}$$

and

$$\alpha_i G_i(\underline{x}) = 0 \tag{7}$$

The α_i and β_i are the dual variables associated with the inequality constraints (3) and the equality constraints (4), respectively. Condition (7) implies that $\alpha_i = 0$ if constraint $G_i(\underline{x})$ is not reached and $\alpha_i > 0$ in the alternative.

B. The Network Equations

Under steady-state conditions, a power system of N + 1 nodes is described by the following 2(N + 1) nonlinear algebraic equations

---

[*] References are listed at the end of the paper.

$$P_i - C_i = \sum_{\alpha} \frac{U_i U_\alpha}{Z_{i\alpha}} \sin(\theta_i - \theta_\alpha - \delta_{i\alpha}) + \sum_{\alpha} \frac{U_i^2 \sin \delta_{i\alpha}}{Z_{i\alpha}} \tag{8}$$

$$Q_i - D_i = -\sum_{\alpha} \frac{U_i U_\alpha}{Z_{i\alpha}} \cos(\theta_i - \theta_\alpha - \delta_{i\alpha}) + \sum_{\alpha} \frac{U_i^2 \cos \delta_{i\alpha}}{Z_{i\alpha}} - Y_{ii} U_i^2$$

$$i = 0,1,\ldots,N; \; \alpha = 0,1,\ldots,i-1, i+1,\ldots,N. \tag{9}$$

where

$Z_{i\alpha}$ = magnitude of the impedance between nodes i and α

$\delta_{i\alpha}$ = loss angle of the tie-line $i\alpha$; $\delta_{i\alpha} = \sin^{-1} \frac{R_{i\alpha}}{Z_{i\alpha}}$

Y_{ii} = magnitude of the (reactive) admittance between node i and ground.

A convenient and consistent set of units is megawatts (MW) for P and C, megavars (Mvar) for Q and D, kilovolts (KV) for U, ohms (Ω) for Z, and Ω^{-1} for Y.

C. The Equality Constraints

The equality constraints in the problem under discussion state that the difference between the powers produced and consumed at node i, that is $P_i - C_i$ and $Q_i - D_i$, are injected into the system governed by (8) and (9). These "injections" are denoted by I_i and K_i, respectively. Thus:

$$I_i - P_i + C_i = 0 \tag{10}$$

$$K_i - Q_i + D_i = 0 \tag{11}$$

The dual variables associated with (10) and (11) are λ_i and μ_i.

D. The Inequality Constraints

In the general case, there are ten inequality constraints per node. Denoting by the superscripts M and m the upper and lower bounds, these inequality constraints and their associated dual variables are

$P_i^2 + Q_i^2 - S_i^{2,M} \leq 0$ (dual variable M_i) (12)

$P_i^m - P_i \leq 0$ (" " m_i) (13)

$Q_i - Q_i^M \leq 0$ (" " e_i) (14)

$Q_i^m - Q_i \leq 0$ (" " e_i') (15)

$U_i - U_i^M \leq 0$ (" " u_i) (16)

$U_i^m - U_i \leq 0$ (" " u_i') (17)

The constraints (12), (13), (14) and (15) are primarily determined by current magnitudes, boiler instability, generator excitation limits, generator instability, insulation and transformer tap ranges, respectively.

The inequality constraints affecting each pair of nodes are

$$\theta_i - \theta_\alpha - T^M_{i\alpha} \le 0 \qquad \text{(dual variable } t_{i\alpha}\text{)} \tag{18}$$

$$-T^m_{i\alpha} - (\theta_i - \theta_\alpha) \le 0 \; . \tag{19}$$

Inequality constraint (19) is redundant, as one would write for node α

$$\theta_\alpha - \theta_i - T^M_{\alpha i} \le 0 \tag{20}$$

and can therefore be omitted. Angle constraint (19) arises from system stability considerations.

E. The Optimization Equations

The function L of (5) is written as

$$\begin{aligned} L = {} & F(P_0,\ldots,P_N) + \sum_i \lambda_i(I_i - P_i + C_i) + \sum_i \mu_i(K_i - Q_I + D_i) \\ & + \sum_i M_i\,(P_i^2 + Q_i^2 - S_i^{2,M}) + \sum_i m_i\,(P_i^m - P_i) + \sum_i e_i(Q_i - Q_i^M) \\ & + \sum_i e'_i\,(Q_i^m - Q_i) + \sum_i u_i(U_i - U_i^M) + \sum_i u'_i\,(U_i^m - U_i) \\ & + \sum_{i\alpha} t_{i\alpha}\,(\theta_i - \theta_\alpha) \end{aligned} \tag{21}$$

$$i = 0,\ldots,N;\ \alpha = 0,\ldots,i-1,\ i+1,\ldots,N.$$

Setting $dL = 0$, the following group of optimization equations results for a node with production and consumption.

1. Differentiation with Respect to P_i and Q_i

$$\lambda_i = \frac{\partial F}{\partial P_i} + 2M_i\,P_i - m_i \tag{A}$$

$$\mu_i = 2M_i\,Q_i + e_i - e'_i \tag{A'}$$

2. Differentiation with Respect to θ_i and U_i

$$\lambda_i \frac{\partial I_i}{\partial \theta_i} + \sum_\alpha \lambda_\alpha \frac{\partial I_\alpha}{\partial \theta_i} + \mu_i \frac{\partial K_i}{\partial \theta_i} + \sum_\alpha \mu_\alpha \frac{\partial K_\alpha}{\partial \theta_i} + \sum_\alpha (t_{i\alpha} - t_{\alpha i}) = 0 \tag{B}$$

$$\lambda_i \frac{\partial I_i}{\partial U_i} + \sum_\alpha \lambda_\alpha \frac{\partial I_\alpha}{\partial U_i} + \mu_i \frac{\partial K_i}{\partial U_i} + \sum_\alpha \mu_\alpha \frac{\partial K_\alpha}{\partial U_i} + u_i - u'_i = 0 \tag{B'}$$

3. Injection Equations

$$I_i - P_i + C_i = 0 \tag{C}$$

$$K_i - Q_i + D_i = 0 \tag{C'}$$

4. Exclusion Equations

$$M_i(P_i^2 + Q_i^2 - S_i^{2,M}) = 0 \qquad m_i(P_i^m - P_i) = 0$$

$$e_i(Q_i - Q_i^M) = 0 \qquad e'_i(Q_i^m - Q_i) = 0$$

$$u_i(U_i - U_i^M) = 0 \qquad\qquad u'_i(U_i^m - U_i) = 0$$

$$t_{i\alpha}(\theta_i - \theta_\alpha - T_{i\alpha}) = 0 \qquad\qquad \text{(D)}$$

If, at a particular node, there is no production ($P_i = 0$ or $Q_i = 0$ or both), the corresponding A or A′ (or both) equations disappear. In all cases, the number of equations equals the number of variables. If no constraint is met, the variables are P_i, Q_i, λ_i, μ_i, θ_i, U_i; after a constraint has been met, one of these "primary" variables becomes known and the corresponding dual variable takes its place as an unknown.

F. The Significance of the Dual Variables

Since at the optimum dL = 0, it follows from (21) that

$$\begin{aligned} dF = & -\sum_i \lambda_i\, d(I_i - P_i + C_i) - \sum_i \mu_i\, d(K_i - Q_i + D_i) \\ & - \sum_i M_i\, d(P_i^2 + Q_i^2 - S_i^{2,M}) - \sum_i m_i\, d(P_i^m - P_i) - \sum_i e_i\, d(Q_i - Q_i^M) \\ & - \sum_i e'_i\, d(Q_i^m - Q_i) - \sum u_i\, d(U_i - U_i^M) - \sum_i u'_i\, d(U_i^m - U_i) \\ & - \sum_{i\alpha} t_{i\alpha}\, d(\theta_i - \theta_\alpha)\ . \end{aligned} \qquad (22)$$

Suppose that a small change in the minimum power, dP_i^m, were allowed; then, from (21)

$$m_i = \frac{dF}{dP_i^m} \qquad ; \qquad (23)$$

that is, m_i measures the sensitivity of F with respect to the inequality constraint P_i^m at the optimum. Similarly, M_i, e_i, e'_i, u_i, u'_i and $t_{i\alpha}$ measure the sensitivity of F with respect to $S_i^{2,M}$, Q_i^m, U_i^M, U_i^m and $T_{i\alpha}$. The dual variables consequently point out the equipment bottlenecks and thus aid in short-term system planning.

By invoking similar arguments, it is shown that

$$\lambda_i = \frac{dF}{dC_i} \quad ; \qquad \mu_i = \frac{dF}{dD_i} \quad ; \qquad (24)$$

that is, the dual variables λ_i and μ_i; associated with the equality constraints measure the sensitivity of cost with respect to consumption of active and reactive power. They hence provide a rational basis for tarification.

G. Computational Approach

The optimization equations (A), (A′), (B), (B′), (C) and (C′) together with the exclusion equation (D) constitute a set of simultaneous nonlinear algebraic equations, the machine solution of which is possible for systems comprising up to several hundred nodes[1,2].

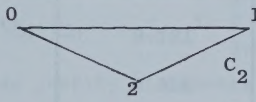
From the point of view of computations, it is best to resolve this set by Newton's method applied to each node in succession. If, during the iteration process, a variable exceeds its admissible range, provision is made to give it the corresponding upper or lower constraint value and to satisfy the linearized equations by forcing the correction into another as yet unconstrained variable. In view of these changes, the establishment of a efficient and reliable program is fairly laborious.

H. Example

A three-node system, as shown in Fig. 1, was treated for the following numerical data:

$C_0 = 400$ MW $P_0 = ?$ $U_0 = ?$ $P_1 = ?$ $U_1 = ?$ $C_1 = 400$

$D_0 = 200$ Mvar $Q_0 = ?$ $\theta_0 = 0$ $Q_1 = ?$ $\theta_1 = ?$ $D_1 = 200$



$C_2 = 800$ $P_2 = 0$ $U_2 = ?$

$D_2 = -337$ $Q_2 = ?$ $\theta_2 = ?$

Fig. 1: Three node system example with $Z_{i\alpha}=50.176\Omega$, $\delta_{i\alpha} = 0.25$ rad, $F = 3P_0 + 10^{-6}P_0^3 + 2P_1 + \frac{2}{3}10^{-6}P_1^3$ \$/hour $200 \le U_i \le 240$ KV.

In Table I below, the computer printout of the results of the first and sixth iteration summarizing the partial iterations performed in sequence at each node is given. Convergence is obtained after six iterations; that is less than 5 sec on a CDC 3100.

The very small value of μ_2 indicates that for $D_2 = -337$Mvar, reactive production at Node 2 is unnecessary. Nodes 1 and 2 are operated at U^M. Node 2 is at a lower voltage, as dictated by D_2. The values of λ indicate that the cost of power should be highest at Node 2, which is reasonable since there is no production P_2.

In Ref. 1, the same numerical example is treated for fixed voltage $U_i = 224$ KV. The resulting minimum cost is 5125.2 \$/hour, as compared to 5055 \$/hour; the transmission losses for both cases are 105.7 MW and 89.4 MW.

II -- LONG-TERM POWER SYSTEM PLANNING

In Part I, the instantaneous optimization of power flow was discussed

Variables and Dual Variables	Iteration No. 1	Iteration No. 6
θ_1	.0719	.1066
θ_2	-.3194	-.3388
U_0	250.3	240.0
U_1	262.3	240.0
U_2	238.4	235.0
λ_0	4.41	4.33
λ_1	4.09	4.09
λ_2	5.14	5.20
P_0	686.4	667.0
P_1	1022.3	1022.4
Q_0	191.8	228.3
Q_1	336.0	184.8
u_0	0	1.27
u_1	0	2.59
u_2	0	0
μ_2	0.1192	-0.0001
F	5139.1689	5055.0031

and the role of the dual variables for short-term planning and tarification was pointed out. In Part II, the problem of optimally scheduling plant investment over the planning interval (ten to twenty years) to meet predicted demand is considered. It is shown that long-term planning is an optimum control problem, which, in the first approximation, can be handled by linear programming; for more realistic assumptions, nonlinear programming and the Kuhn and Tucker conditions apply.

A. Problem Formulation

Given:

(1) The present generating and transmission capability (production capability) denoted by the vector $\underline{x}_0$.

(2) The demand $\underline{d}_k$ at time k (k = 1,...,N) predicted over the planning interval [0,N].

(3) The relation between the production capability $\underline{x}_k$ and the decision to invest $\underline{u}_k$

$$\underline{x}_{k+1} - \underline{f}(\underline{x}_k, \underline{u}_k, k) = 0$$

$$\text{or} \qquad \underline{H}_k(\underline{x}_{k+1}, \underline{x}_k, \underline{u}_k, k) = 0 \tag{24}$$

(4) The inequality constraints between production capability, $\underline{x}_k$, investment decision $\underline{u}_k$, and demand $\underline{d}_k$

$$\underline{G}_k(\underline{x}_k, \underline{u}_k, k) \leq 0 \tag{25}$$

(5) The terminal constraints

$$\underline{S}(\underline{x}_N) \leq 0 \tag{26}$$

(6) The net cost of operating during the interval [k, k+1]

$$\ell_k(\underline{x}_k, \underline{d}_k, \underline{h}_k, k) \tag{27}$$

Find: The investment schedule $\underline{u}_0, \ldots, \underline{u}_{N-1}$ that minimizes the total cost incurred over the planning interval

$$J(\underline{x}_0; \underline{u}_0, \ldots, \underline{u}_{N-1}) = \sum_{k=0}^{N-1} \ell_k(\underline{x}_k, \underline{d}_k, \underline{u}_k, k) \tag{28}$$

The following explanations justify statements (1) through (6):

(1) The components of the vector $\underline{x}_k$ are the generating (thermal, hydro, nuclear etc) and transmission capabilities in MW of the system partitioned into a sufficiently fine regional grid. Depending on the accuracy desired, additional parameters characterizing the plant, such as efficiency, age, etc., are included in $\underline{x}_k$.

(2) The predicted demand $\underline{d}_k$, whose components correspond to regional demand as well as sales to adjacent systems, may need to be further categorized into peak and average values; $\underline{d}_k$ is a random variable, since future demand is not known exactly.

(3) The difference equation (24) accounts for the lag between increased production capability and investment decision.

(4) The inequality constraint (25) forces production to be at least equal to demand and maintains the proper balance between the various production categories, e.g., hydro and thermal.

(5) The terminal constraint (26) ensures that the system is in an acceptable state at the end of the planning interval.

(6) The cost per interval $\ell_k(\underline{x}_k, \underline{u}_k, \underline{d}_k, k)$ is the difference between the sum of the operating costs and investment charges on one hand, and the revenue from power sales on the other hand during the interval [k, k+1].

B. Optimization Equations

If Eqs. (24) through (27) are linear in the variables $\underline{x}_k, \underline{u}_k (k=0,\ldots,N)$, the resulting optimization problem is best resolved by well-known linear programming techniques. The upper bound on the number of variables is of the order of 1000,e.g., 50 state and decision variables for a 20-year planning interval with yearly decision intervals. The results obtained by linear programming are reported in Ref. 5.

In the more realistic case of nonlinear Eqs. (24) through (27), the Kuhn and Tucker conditions apply as follows.

$$L(\underline{x}_1,\ldots,\underline{x}_N;\ \underline{u}_0,\ldots,\underline{u}_{N-1}) \overset{\Delta}{=} \sum_{k=0}^{N-1} \ell_k(\underline{x}_k,\underline{u}_k,k)$$

$$+\sum_{k=1}^{N-1} \underline{\lambda}_k^T\ \underline{H}_k(\underline{x}_{k+1},\ \underline{x}_k,\underline{u}_k,k) + \sum_{k=0}^{N-1} \underline{\mu}_k^T\ \underline{G}_k(\underline{x}_k,\underline{u}_k,k) + \underline{\nu}^T\ \underline{S}(\underline{x}_N) \quad (29)$$

Subject to satisfaction of convexity conditions on L, the optimization equations are obtained as before:

1. Differentiation with Respect to $\underline{x}_k$ (k=1,...,N-1)

$$\nabla_{k_k} \{\ell_k + \underline{\lambda}_k^T\ \underline{H}_k + \underline{\lambda}_{k-1}^T\ \underline{H}_{k-1} + \underline{\mu}_k^T\ \underline{G}_k\} = 0 \quad (30)$$

2. Differentiation with Respect to u_k(k=0,...,N-1)

$$\nabla_{u_k} \{\ell_k + \underline{\lambda}_k^T\ \underline{H}_k + \underline{\mu}_k^T\ \underline{G}_k\} = 0 \quad (31)$$

3. Differentiation with Respect to $\underline{x}_N$.

$$\nabla_{x_N} \{\underline{\lambda}_{N-1}^T\ \underline{H}_{N-1} + \underline{\nu}^T \underline{S}\} = 0 \quad (32)$$

To the optimization equations (30), (31) and (32), the exclusion equations

$$\underline{\mu}_k^T\ \underline{G}_k\ (\underline{x}_k,\underline{u}_k,k) = 0 \qquad k=0,\ldots,N-1 \quad (33)$$

$$\underline{\nu}^T\ \underline{S}(x_N) = 0 \quad (34)$$

and the equality constraint (24) must be added.

The number of unknown vectors is N for $\underline{x}$, N for $\underline{u}$ and N for λ; the number of sets of optimization equations is N-1 for (30), N for (31), one for (32) and N for (24).

The dual variables $\underline{\mu}_k$ and $\underline{\upsilon}$ associated with the inequality constraints (25) and (26) again measure the sensitivity of the minimum cost $\min\limits_{\underline{u}_0,\ldots,\underline{u}_{N-1}} J$ with respect to constraints.

C. General Remarks

Although the optimum planning problems can always be formulated as in Sec. A and although the Kuhn and Tucker conditions (30) - (34) can always be obtained for differentiable L, the resulting investment schedule is certain to constitute a global optimum only if L is convex.

After the Kuhn and Tucker conditions (30) - (34) have been obtained, the machine solution of the resulting 3N sets of simultaneous nonlinear equations will in general require a major programming effort and may not be solvable at all depending on the number of variables and the form of the equations (24), (25), (26), and (27). Despite these difficulties, the method is capable of solving planning problems of a very respectable size (up to a thousand variables under favorable conditions) and in addition provides quantitative expressions for the sensitivity of cost with respect to the constraints imposed.

It was recently shown[4] that the optimization conditions of Kuhn and Tucker (30)-(32), which constitute difference equations in $\underline{x}, \underline{u}, \underline{\lambda}$ and $\underline{\upsilon}$ reduce to the optimization equations derived from Pontryagin's maximum principle when the decision interval $[k, k+1]$ tends toward zero.

The demand $\underline{d}_k$ is a random variable, as are the prices and operating costs of the various production categories, with the result that the problem of power system planning actually becomes a stochastic optimum control problem. To bypass the resulting additional difficulties, it is convenient and realistic to formulate a deterministic optimization problem for a "worst case," an average and a "best case" hypothesis and to determine thereafter a planning schedule which is optimum for the hypothesis adopted.

REFERENCES

1. Carpentier, J., "Contribution à l'Etude du Dispatching Economique," *Bulletin de la Société Francaise des Electriciens*, Ser. 8, Vol. 3, (August, 1962).
2. Carpentier, J., and J. Sirous, "L'optimisation de la Production à l'Electricité de France," *Bulletin de la Société Francaise des Electriciens* (March 1963).
3. Kuhn, H.W. and A.W. Tucker, "Nonlinear Programming" *Proc. of 2nd Berkeley Symposium on Math.Stat. and Prob.*, (University of Cal. Press, Berkeley, California, 1951).
4. Mangasarian, O.L., "Sufficient Conditions for the Optimal Control of Nonlinear Systems," presented at the First International Conference on Programming and Control, Colorado Springs, Colorado, April 15-16, 1965; to appear in the SIAM Journal.
5. Gourssens, P., *et al.* "The Application of Digital Computers to Power Transmission Network Studies and Operation," *Conference Internationale des Grands Réseaux Electriques à Haute Tension*, 1-10 Jun, 1964, Paris.

OPTIMAL TIME-DOMAIN SYNTHESIS BY ITERATIVE METHODS

L. E. McBRIDE, H. W. SCHAEFGEN[*] and K. STEIGLITZ
Department of Electrical Engineering, Princeton University,
Princeton, New Jersey

ABSTRACT

An iterative procedure is presented which permits the determination of a rational transfer function in the Laplace transform variable s which is optimal with respect to given input and output time-functions. The optimal system of a particular order is defined as the one whose output when subjected to the known input function is nearest in the time integral square sense to the desired output function. The method is thus applicable to a number of problems involving the minimization of an integral square error.

To illustrate the technique, a set of optimal lumped-parameter delay lines is synthesized and their characteristics investigated; the behavior and convergence of the iteration in these problems is also studied.

---

[*] Now with Aeronutronic Division, Ford Motor Co., Newport Beach, California.

PROBLEM FORMULATION

Time-domain approximation can be regarded as the problem of finding a network of given complexity which, when subjected to a particular input, will produce (at least approximately) a desired output as a function of time. If it is assumed that any rational function of given order can be synthesized as the transfer function of a network, the problem reduces to finding the coefficients of a ratio of polynomials in the transform variable s which best fits the required input-output relation. It is in this sense identical to the problem of identifying an unknown linear system from a sample of its input and output functions.

Among the computational procedures proposed for linear system identification, Kalman's [1] method for finding the coefficients of a rational z-transform function by linear regression is among the easiest to use. However, as a method for time-domain synthesis it has two disadvantages: it results in a sampled-data model which must then be approximated by a rational function of s, and it minimizes an artificial error criterion which does not correspond directly to any measurable error. It has been shown [2] that the second objection can be eliminated by using an iterative procedure, with the Kalman estimate as first iteration; if the iteration converges, as it does in many practical cases, the resulting transfer function minimizes the mean square error between desired and actual model output. It is the purpose of this report to show that similar iterative procedures can be applied directly to continuous system identification and to give examples of their effectiveness in time-domain approximation.

Let the system input be x(t) and the desired output function y(t), both defined for $t \geq 0$ (x(t) may be an ideal impulse, in which case y(t) is the desired network impulse response). If v(t) is the actual response of the network to x(t), a natural error criterion is

$$J = \int_0^T [v(t) - y(t)]^2 dt = \int_0^T e^2(t)\, dt \tag{1}$$

The transfer function of the network to be synthesized has the form

$$\begin{aligned} G(s) &= \frac{a_n + a_{n-1} s + a_{n-2} s^2 + \ldots + a_o s^n}{b_n + b_{n-1} s + b_{n-2} s^2 + \ldots + s^n} \\ &= \frac{a_n s^{-n} + a_{n-1} s^{-n+1} + \ldots + a_o}{b_n s^{-n} + b_{n-1} s^{-n+1} + \ldots + 1} = \frac{N(s)}{D(s)} \end{aligned} \tag{2}$$

(In most problems a_o is omitted because it is known that the network impulse response should not contain an impulse at t = 0; it is also possible to assume a numerator order different from that of the denominator, but this alternative will be ignored to simplify the notation.) Then if X(s) is the Laplace transform of x(t), etc.,

$$V(s) = \frac{N(s)}{D(s)} X(s) \tag{3}$$

and

$$\begin{aligned} \frac{\partial E(s)}{\partial a_i} &= \frac{s^{-i}}{D(s)} X(s) = P_{ai}(s) \\ \frac{\partial E(s)}{\partial b_i} &= \frac{-s^{-i} N(s)}{D^2(s)} X(s) = -P_{bi}(s) \end{aligned} \tag{4}$$

Now

$$\begin{aligned} \frac{\partial J}{\partial a_i} &= 2 \int_0^T e(t) \frac{\partial e(t)}{\partial a_i}\, dt \\ \frac{\partial J}{\partial b_i} &= 2 \int_0^T e(t) \frac{\partial e(t)}{\partial b_i}\, dt \end{aligned} \tag{5}$$

and, by the definition of the Laplace transform,

$$P_{ai}(s) = \frac{\partial}{\partial a_i} \int_0^\infty e(t) \exp(-st) dt = \int_0^\infty \frac{\partial e(t)}{\partial a_i} \exp(-st)\, dt; \tag{6}$$

therefore:

$$p_{ai}(t) = \frac{\partial e(t)}{\partial a_i}, \; p_{bi}(t) = \frac{-\partial e(t)}{\partial b_i},$$

and a necessary condition for the values of a_i and b_i to be optimal with respect to the error criterion J is

$$\begin{aligned} &\int_0^T e(t)\, p_{ai}(t)\, dt = 0 \qquad i = 1, \ldots n \\ &\int_0^T e(t)\, p_{bi}(t)\, dt = 0 \qquad i = 1, \ldots n \end{aligned} \tag{7}$$

where p_{ai}, p_{bi} may be regarded as the output from a set of filters described by (4).

Unfortunately, (7) is a highly nonlinear equation in the coefficients a_i and b_i, and no method of direct solution is apparent. Three ways of attacking the problem have been considered; they can be called for convenience Gradient Methods, Mode 1 Iteration, and Mode 2 Iteration.

Gradient Methods

Since the left-hand side of (7) is proportional to the gradient of J in coefficient space, one straightforward computational method is based on a descent to the minimum. For any initial set of a's and b's, e(t), $p_{ai}(t)$, and $p_{bi}(t)$ can be obtained by simulating the filtering operations of (3) and (4) on either a digital or analog computer. The gradient can then be computed from (7) and indicates in which direction each coefficient must be altered to reduce J. A suitable small increment is then added to each coefficient, the resulting e, p's and gradient computed and the process repeated until the gradient is approximately zero. A variety of more sophisticated programming methods for finding the minimum given a gradient computation procedure is also available[3].

These methods have been tested on similar problems in sampled-data systems and have been found effective. Usually, however, a very large number of iterations is required for convergence, and the procedure is particularly unwieldy when a large number of coefficients (e.g., more than ten) is to be computed; because the other iterative methods seem to work so well for time domain synthesis, gradient techniques were not tried in this case.

Mode 1 Iteration

A new error, $e_1(t)$, is defined such that

$$E_1(s) = \frac{N_j(s)}{D_{j-1}(s)} X(s) - \frac{D_j(s)}{D_{j-1}(s)} Y(s) \tag{8}$$

where the subscript j refers to the iteration number. It is assumed that the coefficients of D_{j-1} have been previously computed, and are regarded as fixed during the jth iteration. Therefore,

$$\begin{aligned} \frac{\partial E_1(s)}{\partial a_i} &= \frac{s^{-1}}{D_{j-1}(s)} X(s) = \widehat{P}_{ai}(s) \\ \frac{\partial E_1(s)}{\partial b_i} &= \frac{-s^{-i}}{D_{j-1}(s)} Y(s) = -\widehat{P}_{bi}(s) \end{aligned} \tag{9}$$

Since the expressions in (9) are independent of a_i and b_i, and $e_1(t)$ is a linear function of these coefficients, the gradient of J is linear and the optimal values of the coefficients can be obtained by linear regression.

If the coefficient vector $\underline{\gamma}$ and signal vector $\underline{q}(t)$ are defined by

$$\underline{\gamma} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ -b_1 \\ \vdots \\ -b_n \end{bmatrix} \qquad \underline{q}(t) = \begin{bmatrix} \hat{p}_{a1}(t) \\ \vdots \\ \hat{p}_{an}(t) \\ \hat{p}_{b1}(t) \\ \vdots \\ \hat{p}_{bn}(t) \end{bmatrix},$$

$e_1(t)$ and J_1 can be written as

$$\begin{aligned} e_1(t) &= \underline{\gamma}'\,\underline{q}(t) - \hat{p}_{bo}(t) \\ J_1 &= \int_0^T e_1^2(t)\,dt \end{aligned} \tag{10}$$

and the gradient of J_1 with respect to $\underline{\gamma}$ is

$$\begin{aligned} \text{grad } J_1 &= 2 \left[\int_0^T \underline{q}(t)\,\underline{q}'(t)\,dt \right] \underline{\gamma} - 2 \int_0^T \hat{p}_{bo}(t)\,\underline{q}(t)\,dt \\ &= 2\underline{Q}\,\underline{\gamma} - 2\underline{c} \end{aligned} \tag{11}$$

where the prime denotes the transpose, $\underline{Q}$ is a 2n x 2n matrix and $\underline{c}$ is a vector of order 2n.

Then the values of the 2n a's and b's which minimize the integral square error J_1 can be obtained directly from

$$\underline{\gamma} = \underline{Q}^{-1}\underline{c} \tag{12}$$

Since all of the signals composing $\underline{q}(t)$ and $\hat{p}_{bo}(t)$ can be obtained from the given x(t) and y(t) by means of the filtering defined by (9), $\underline{Q}$ and $\underline{c}$ can be computed either by analog or digital simulation from the previously estimated denominator $D_{j-1}(s)$.

The difference between the true error e(t) and the "linear error" $e_1(t)$ is shown graphically in Fig. 1. If $D_o(s)$ is assumed to be unity, e_1 is simply the true error filtered by a device having $D_1(s)$ as its transfer function. Therefore in the first iteration e_1 is analogous to the Kalman error used in sampled-data system identification [1, 2].

The iteration can thus be started either with a first estimate of D obtained from some other source, or with the equivalent of a Kalman estimate. The computation of $\underline{\gamma}$ is repeated; if at some stage in the iteration $D_j(s) \approx D_{j-1}(s)$, then $e_1(t) \approx e(t)$ and $J_1 \approx J$. Such convergence cannot be guaranteed, but experimental evidence indicates that the procedure does converge in many practical problems, and produces a significant reduction in true integral square error compared with the Kalman equivalent or various non-optimal approximation techniques.[1]

Mode 2 Iteration

When the Mode 1 iteration converges, $J_1 = J$ but grad $J_1 \neq$ grad J. Since it is the gradient that enters into the condition for a minimum, it might be expected that a procedure converging to grad J = 0 would reduce the error still further.

Let a new signal variable $\tilde{p}_{bi}(t)$ be defined by its transform

$$\tilde{P}_{bi}(s) = \frac{s^{-i} N_{j-1}(s)}{D^2_{j-1}(s)} X(s) = \frac{s^{-i}}{D_{j-1}(s)} V(s). \tag{13}$$

Then an approximation to (5) can be written

$$\begin{aligned} \frac{\partial J}{\partial a_i} &\approx 2 \int_0^T e_1(t)\, \hat{p}_{ai}(t)\, dt \\ \frac{\partial J}{\partial b_i} &\approx -2 \int_0^T e_1(t)\, \tilde{p}_{bi}(t)\, dt \end{aligned} \tag{14}$$

and since the right-hand side of (14) is linear in a_i, b_i, it can be set equal to zero and solved for the coefficients.

The computation is similar to that of Mode 1. If a signal vector $\underline{p}(t)$ is defined by

---

[1] A similarly motivated iteration, proposed by Loeb [6] for fitting a ratio of polynomials to a known function in the Tchebycheff sense, has been found strongly convergent in such problems [7].

$$\underline{p}'(t) = \left[\hat{p}_{a1}(t)\dots\hat{p}_{an}(t)\ \ \tilde{p}_{b1}(t)\dots\tilde{p}_{bn}(t)\right] \tag{15}$$

the gradient of J with respect to $\underline{\gamma}$ becomes approximately

$$\begin{aligned} \text{grad } J &\approx 2\left[\int_0^T \underline{p}(t)\,\underline{q}'(t)\,dt\right]\underline{\gamma} - 2\int_0^T p_{bo}(t)\,\underline{p}(t)\,dt \\ &= 2\underline{\tilde{Q}}\underline{\gamma} - 2\,\underline{c} \end{aligned} \tag{16}$$

It should be noted that this mode cannot be started from an initial estimate of G(s) which is either zero or unity. In the first case, $\tilde{p}_{bi} = 0$, and in the second $\tilde{p}_{bi} = \hat{p}_{ai}$; in either case $\underline{\tilde{Q}}$ will be singular. At least one estimate must be made by a different method before Mode 2 can be used. As might be expected, the convergence of Mode 2 is more variable than that of Mode 1, but when it converges rapidly it attains the true minimum.

Because Mode 1 uses the desired output y(t) as a substitute for the output v(t) from the optimal N/D (compare (9b) and (13)), it seems reasonable that in the early stages of iteration when v(t) is far from optimal, Mode 1 should give a better approximation to the true error surface, and therefore converge more rapidly. On the other hand, once the solution is in the neighborhood of the optimal point, the actual v is the best approximation to the optimal v and Mode 2 converges more rapidly. This general expectation has been borne out by experience in every case considered.

SIMULATION RESULTS

The two iterative techniques were applied to the problem of approximating the unit step response of a lumped-parameter transfer function to that of an ideal delay line. The delay time was one second, and the error squared was integrated from t = 0 to t = 6 sec. to form the quality criterion (the result approximates that which would be obtained over a semi-infinite interval, since the error is very small beyond t = 4). A summary of the error criterion values resulting from various approximations is given in Table 1. The initial parameter values from which the iterations were started were either a) all zeroes, in which case the first iteration yields the Kalman-equivalent estimate, or b) coefficients obtained from the maximally flat approximation to a time delay [4, 5] taken from reference 5. A commonly applied figure of merit, the delay-to-rise time ratio, is also given for the final set of parameters in each run (rise time is defined as the interval during which the response is between 5% and 95% of its final value). As a standard for comparison, this ratio for the eighth-order maximally flat network is approximately 1.2.

The step responses of the optimal third, fifth, and eighth-order networks from Mode 2 are shown in Figs. 2, 3, and 4. In Fig. 5 the response of the fifth-order Mode 1 network is shown for comparison. Comparison of Figs. 3 and 5 demonstrates the fact that although the reduction in error by the use of Mode 2 is only about 4%, the resulting step response is clearly superior.

Reduction of integral square error by the iteration process is illustrated in Fig. 6. The errors shown for iteration 0 are those of the maximally flat approximation (off-scale in second- and third-order cases). The slightly erratic behavior for $N = 7$ was found to be caused by too large a step size in the numerical integration routine. Note, however, that Mode 1 exhibits a relative minimum at the second or third iteration in nearly every case; such behavior seems to be typical of this procedure.

Figure 7 shows the manner in which the poles of the eighth-order network are shifted from their initial values to their final locations by tye iterative procedures. Changes after the fourth iteration in Mode 1 (solid arrows) are too small to be effectively shown on this scale; it is clear that the requirement that $e_1 \approx e$ is fulfilled after very few iterations. (The initial poles shown are the result of one iteration started from the maximally flat approximation.) The dashed arrows show the effect of two iterations in Mode 2, starting from the final Mode 1 poles.

CONCLUSION

A method has been proposed which is applicable to the general problem of optimal time-domain synthesis, defined as the determination of the rational transfer function of given order whose output when subjected to a known input function best approximates, in a mean square sense, a given desired time function. Since the problem of linear system identification from operating records may be posed in similar fashion, the technique is applicable to both.

As an illustration of the practicality of the iterative approach, lumped delay lines of orders from second through eighth have been synthesized; they are optimal for step inputs. Effective convergence in such problems appears to occur within three to five iterations. The method appears to have advantages over other iterative procedures for this class of problems, requiring fewer iterations for convergence and simpler computations for each iteration. The proposed technique is easily programmed for the digital computer, but requires longer computing times than the equivalent sampled-data identification problem.

Acknowledgment

This work was partially supported by National Science Foundation Grant N.S.F. GK-270 and Army Research Office Contract DA-31-124-ARO-D-292; it made use of computer facilities supported in part by Grant NSF GP-579.

REFERENCES

1. R. E. Kalman, "Design of a Self-Optimizing System," Trans. ASME, vol. 80, pp. 468-78; February, 1958.

2. K. Steiglitz and L. E. McBride, "A Technique for the Identification of Linear Systems," IEEE Trans. on Auto. Control, vol. AC-10; October, 1965.

3. P. J. Woodrow, "Methods for Locating the Extrema of Multivariate Functions," Control Systems Laboratory, Tech. Rept. No. 4, Princeton University.

4. L. Storch, "Synthesis of Constant-Time Delay Ladder Networks Using Bessel Polynomials," Proc. IRE, vol. 42, pp. 1666-1675; 1954.

5. M. E. Van Valkenburg, Introduction to Modern Network Synthesis, John Wiley & Sons, Inc., New York; 1960.

6. H. L. Loeb, A Note on Rational Function Approximation, Convair Astronautics Applied Mathematics Series No. 27, September, 1959.

7. E. W. Cheney and T. H. Southard, "A Survey of Methods for Rational Approximation," SIAM Review, vol. 5, no. 3, pp. 219-231; July 1963.

TABLE 1. Errors are 6 Second Integrals of the Error Squared

Model Order	Method	Initial Parameters	Error - Maximally Flat Approximation	Error - [*] Iteration 1	Error - Iteration 2	Error - Final	Delay ÷ Rise Time (final)
2	1	m.f.	.15017	.09382	.09360	.09382	1.1
2	1	0		.09543			
2	2	m.f.	.15017	.09299	.09222	.09171	1.1
3	1	m.f.	.11639	.05644	.05581	.05640	1.9
3	1	0		.05733			
3	2	m.f.	.11639	.05728	.05516	.05459	1.9
4	1	m.f.	.09722	.04040	.03908	.03972	2.6
4	1	0		.06767	.03959	.03927	
4	2	m.f.	.09722	.04143	.03896	.03801	2.6
5	1	m.f.	.08469	.03206	.03004	.03045	3.3
5	1	0		.06711	.03320	.03001	
5	2	m.f.	.08469	.03258	.03020	.02886	3.3
6	1	m.f.	.07578	.02860	.02479	.02462	4.3
6	1	0		.12988	.07010	.02885[**]	
6	2	m.f.	.07579	.03063	.02450	.02314	4.2
7	2	m.f.	.06910	.02817	.02133	.01975	5.0
8	1	m.f.	.06387	.03223	.02034	.01772	6.0
8	2	m.f.	.06387	.02395	.02186	.01678[***]	5.8

---

[*] Error - Iteration 1 gives the error of the Kalman equivalent estimate in the case where I. C. = 0.

[**] Third estimate.

[***] Two iterations of Mode 2 started at the final coefficients obtained from Mode 1.

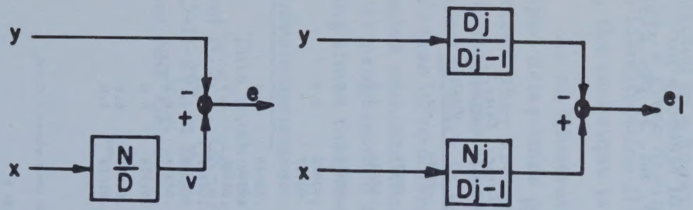

y
x
N
D
v
−
+
e
y
x
Dj
Dj−1
Nj
Dj−1
−
+
e1

FIG. 1.

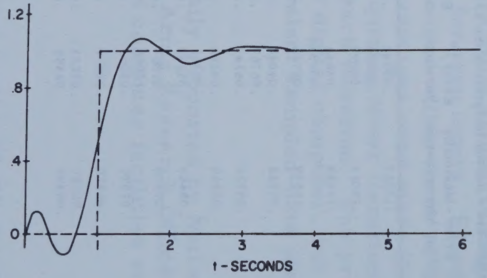


FIG. 2

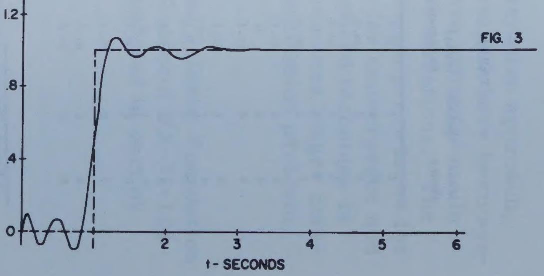


FIG. 3

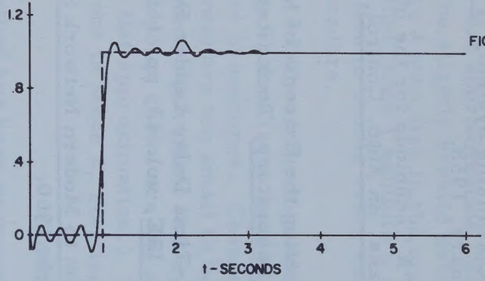


FIG. 4

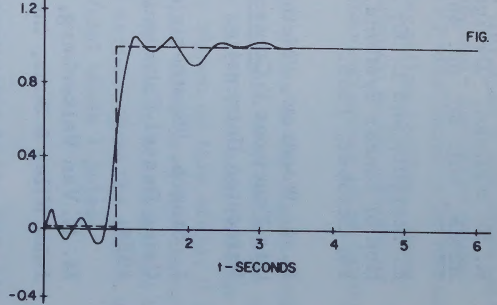


FIG. 5

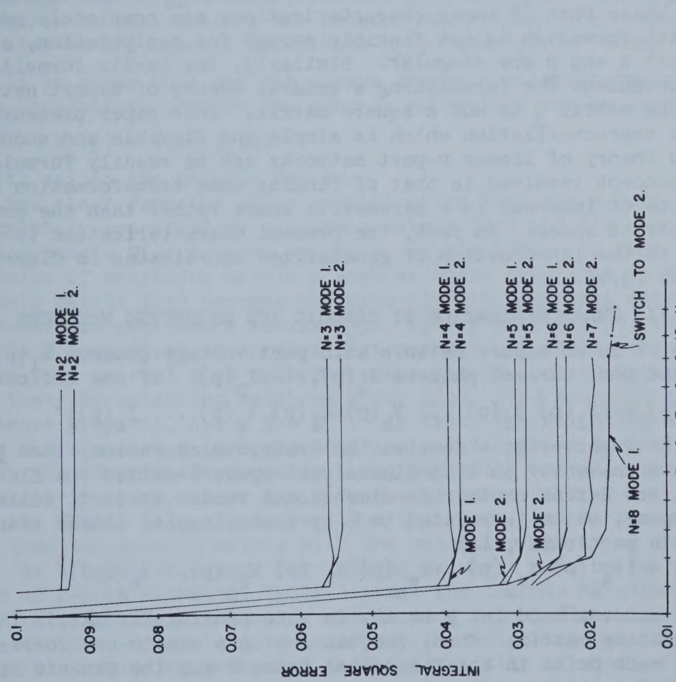

INTEGRAL SQUARE ERROR
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
ITERATION NUMBER
0
1
2
3
4
5
6
7
8
9
10
11
N=2 MODE 1.
N=2 MODE 2.
N=3 MODE 1.
N=3 MODE 2.
N=4 MODE 1.
N=4 MODE 2.
N=5 MODE 1.
N=5 MODE 2.
N=6 MODE 1.
N=6 MODE 2.
N=7 MODE 2.
N=8 MODE 1.
MODE 1.
MODE 2.
MODE 2.
SWITCH TO MODE 2.

FIG. 6.

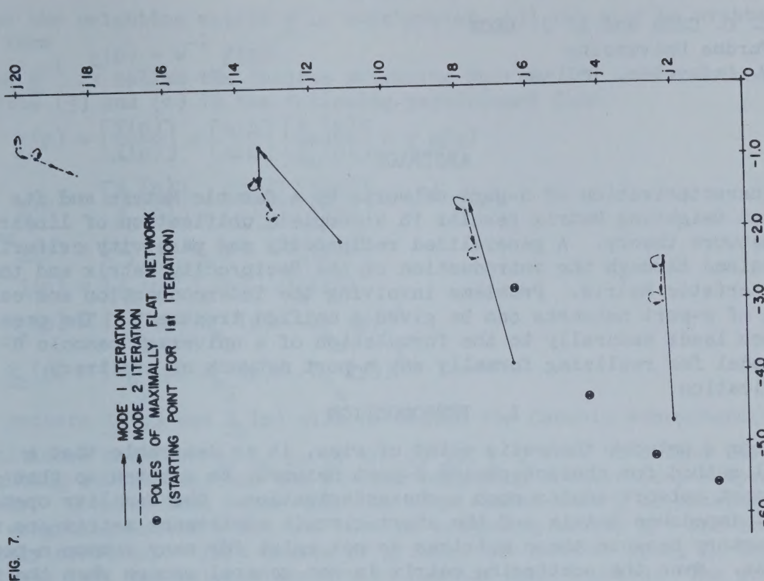

MODE 1 ITERATION
MODE 2 ITERATION
● POLES OF MAXIMALLY FLAT NETWORK (STARTING POINT FOR 1st ITERATION)
j20
j18
j16
j14
j12
j10
j8
j6
j4
j2
0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0

FIG. 7.

A UNIFIED THEORY OF LINEAR N-PORT NETWORKS

L. O. CHUA and B. J. LEON
Purdue University
W. Lafayette, Indiana

ABSTRACT

Characterization of n-port networks by a Canonic Matrix and its associated Weighting Matrix results in a complete unification of linear n-port network theory. A generalized reciprocity and passivity criterion is obtained through the introduction of the Reciprocity Matrix and the Characteristic Matrix. Problems involving the interconnection and cascading of n-port networks can be given a unified treatment. The present approach leads naturally to the formulation of a universal canonic n-port model for realizing formally any n-port network of arbitrary specification.

I. INTRODUCTION

From a network theoretic point of view, it is desirable that a general method for characterizing n-port networks be adopted so that any n-port network admits such a characterization. The familiar open-circuit impedance matrix and the short-circuit admittance matrix are unsatisfactory because these matrices do not exist for many common n-port networks. Even the scattering matrix is not general enough when the network is active. The most general characterizations known today are the $\underset{\sim}{A}$ $\underset{\sim}{B}$ formalism due to Belevitch [1] and the $\underset{\sim}{Q}$ formalism due to Carlin [2]. While both of these characterizations are completely general, the Belevitch formalism is not flexible enough for manipulation, especially when both $\underset{\sim}{A}$ and $\underset{\sim}{B}$ are singular. Similarly, the Carlin formalism is not very convenient for formulating a general theory of n-port networks because the matrix $\underset{\sim}{Q}$ is not a square matrix. This paper presents another general characterization which is simple and flexible and such that a unified theory of linear n-port networks can be readily formulated. The basic concept involved is that of finding some transformation so that the space of interest is a parametric space rather than the usual voltage-current space. In fact, the present characterization is quite analogous to the introduction of generalized coordinates in Classical Dynamics [3,4].

II. CHARACTERIZATION BY CANONIC AND WEIGHTING MATRICES

Let $\mathfrak{N}$ be an n-port network with port voltage phasors $v_1(p),\ldots,$ $V_n(p)$ and port current phasors $I_1(p),\ldots,I_n(p)$. If one defines

$$\underset{\sim}{\xi}(p) \equiv [V_1(p)\ V_2(p)\ \cdots\ V_n(p)\ I_1(p)\ I_2(p)\ \cdots\ I_n(p)]^t \tag{1}$$

where the superscript t denotes the transpose operation, then $\underset{\sim}{\xi}(p)$ represents a 2n-vector in a 2n-dimensional space ϵ-called the Electrical Space. Now define another 2n-dimensional vector space η, called the Canonic Space, which is related to ϵ by a nonsingular linear transformation. In particular, let

$$\underset{\sim}{\varkappa}(p) = [X_{a_1}(p)\ X_{a_2}(p)\cdots X_{a_n}(p)\ X_{b_1}(p)\ X_{b_2}(p)\cdots X_{b_n}(p)]^t \tag{2}$$

be any vector in η, and let $\underset{\sim}{W}$ be any 2n x 2n nonsingular matrix - called the weighting matrix. Then, one can set up a one-to-one correspondence between each point in the Electrical Space ϵ and the Canonic Space η by the transformation

$$\underset{\sim}{\xi}(p) = \underset{\sim}{W}\,\underset{\sim}{\varkappa}(p)\ . \tag{3}$$

Since the weighting matrix $\mathbf{W}$ is nonsignular, (3) can also be written in the form

$$\boldsymbol{\kappa}(p) = \mathbf{W}^{-1}\,\boldsymbol{\xi}(p) \tag{4}$$

where $\mathbf{W}^{-1}$ is called the Inverse Weighting Matrix. It is convenient to rewrite (3) and (4) in the following partitioned form:

$$\boldsymbol{\xi}(p) \equiv \begin{bmatrix} \mathbf{V}(p) \\ \hdashline \mathbf{I}(p) \end{bmatrix} = \begin{bmatrix} \mathbf{a} & \vdots & \mathbf{b} \\ \hdashline \mathbf{c} & \vdots & \mathbf{d} \end{bmatrix} \begin{bmatrix} \mathbf{X}_a(p) \\ \hdashline \mathbf{X}_b(p) \end{bmatrix} \equiv \mathbf{W}\,\boldsymbol{\kappa}(p) \tag{5}$$

and

$$\boldsymbol{\kappa}(p) \equiv \begin{bmatrix} \mathbf{X}_a(p) \\ \hdashline \mathbf{X}_b(p) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha} & \vdots & \boldsymbol{\beta} \\ \hdashline \boldsymbol{\gamma} & \vdots & \boldsymbol{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{V}(p) \\ \hdashline \mathbf{I}(p) \end{bmatrix} \equiv \mathbf{W}^{-1}\,\boldsymbol{\xi}(p) \tag{6}$$

where

$$\mathbf{V}(p) \equiv [V_1(p)\ V_2(p) \ldots V_n(p)]^t \tag{7}$$

$$\mathbf{I}(p) \equiv [I_1(p)\ I_2(p) \ldots I_n(p)]^t \tag{8}$$

$$\mathbf{X}_a(p) \equiv [X_{a_1}(p)\ X_{a_2}(p) \ldots X_{a_n}(p)]^t \tag{9}$$

$$\mathbf{X}_b(p) \equiv [X_{b_1}(p)\ X_{b_2}(p) \ldots X_{b_n}(p)]^t \tag{10}$$

The vectors $\mathbf{X}_a(p)$ and $\mathbf{X}_b(p)$ will be called the Canonic Subvectors; the partitioned matrices $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$, and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ $\boldsymbol{\delta}$ will be called the weighting submatrices and the inverse weighting submatrices, respectively. Observe that other than the requirement that $\mathbf{W}$ be nonsingular, the weighting matrix $\mathbf{W}$ and its inverse weighting matrix $\mathbf{W}^{-1}$ are completely arbitrary. Each choice of $\mathbf{W}$ fixes the canonic space η. Observe that, so far, the network constraint has not been invoked. The presence of an n-port network is to impose some definite constraints on its terminal voltage and current phasor vectors $\mathbf{V}(p)$ and $\mathbf{I}(p)$. Since the weighting matrix is nonsingular, the network must impose a corresponding constraint on the canonic subvectors $\mathbf{X}_a(p)$ and $\mathbf{X}_b(p)$. Since $\mathbf{W}$ is arbitrary, it is clear that one can always choose a weighting matrix $\mathbf{W}$ so that the constraints between the canonic subvectors can be expressed in the form

$$\mathbf{X}_a(p) = \mathbf{T}(p)\,\mathbf{X}_b(p) \tag{11}$$

where $\mathbf{T}(p)$ is called the Canonic Matrix.

It is now clear from the above consideration that any n-port network is completely specified by a canonic matrix $\mathbf{T}(p)$ and its associated weighting matrix $\mathbf{W}$(or inverse weighting matrix $\mathbf{W}^{-1}$). In fact, by a suitable choice of weighting matrix $\mathbf{W}$ such as those tabulated in Table 1, the canonic matrix $\mathbf{T}(p)$ becomes the open-circuit impedance matrix $\mathbf{Z}(p)$, short-circuit admittance matrix $\mathbf{Y}(p)$, hybrid matrix $\mathbf{H}(p)$, scattering matrix $\boldsymbol{\Sigma}$ normalized to complex port numbers and scattering matrix $\mathbf{S}$ normalized to real port numbers, respectively. It is interesting to observe that the weighting matrices $\mathbf{W}$ for $\mathbf{Z}(p)$, $\mathbf{Y}(p)$ and $\mathbf{H}(p)$ exhibit the idempotence property, i.e., $\mathbf{W} = \mathbf{W}^{-1}$. In fact, the weighting matrices $\mathbf{W}$ for the standard 2-port characterization tabulated in Table 2.1 also exhibit the idempotence property. The weighting matrices for other 2-port characterizations, where $\mathbf{W} \neq \mathbf{W}^{-1}$, are given in Table 2.2. Observe that the NULLOR [2], a singular 2-port element, can be characterized by an Identity canonic matrix with the weighting matrix as shown in Table 2.2. As a handy reference for solving networks made up of interconnections of 2-port elements, Table 3 gives the canonic matrices and the associated weighting matrices of some common 2-port elements.

It is sometimes convenient to interpret the weighting matrix $\mathbf{W}$ as playing the role of a coding mechanism which essentially identifies the canonic matrix associated with it. For example, the weighting matrix $\mathbf{W} = \mathbf{1}_n$ identifies the canonic matrix as the open-circuit impedance matrix. One of the nice properties of the present characterization is that it is easy to transform a canonic matrix $T(p)$ and its associated weighting matrix W (or inverse weighting matrix W^{-1}) into another canonic matrix $T'(p)$ with its associated weighting matrix W'^{-1}. In fact, one can readily derive the following useful expressions for transforming $T(p)$ into $T'(p)$:

$$T'(p) = [(\alpha' a+\beta' c)T(p) + (\alpha' b+\beta' d)][(\gamma' a+\delta' c)T(p) + (\gamma' b+\delta' d)]^{-1} \tag{12}$$

$$T'(p) = [(\alpha' a+\beta' c) + (\alpha' b+\beta' d)T^{-1}(p)][(\gamma' a+\delta' c) + (\gamma' b+\delta' d)T^{-1}(p)]^{-1} \tag{13}$$

$$T'(p) = [\alpha'(\alpha-T(p)\gamma)^{-1}(T(p)\delta-\beta)+\beta'][\delta'+\gamma'(\alpha-T(p)\gamma)^{-1}(T(p)\delta-\beta)]^{-1} \tag{14}$$

$$T'(p) = [\alpha'+\beta'(T(p)\delta-\beta)^{-1}(\alpha-T\gamma)][\gamma'+\delta'(T(p)\delta-\beta)^{-1}(\alpha-T(p)\gamma]^{-1} \tag{15}$$

$$T'(p) = [(c\;T(p)+d)(a\;T(p)+b)^{-1}a'-c']^{-1}[d'-(c\;T(p)+d)(a\;T(p)+b)^{-1}b'] \tag{16}$$

$$T'(p) = [(a\;T(p)+b)(c\;T(p)+d)^{-1}c'-a']^{-1}[b'-(a\;T(p)+b)(c\;T(p)+d)^{-1}d'] \tag{17}$$

As an example of the utility of the above equations, suppose it is desired to transform the ABCD 2-port parameters into the open-circuit impedance matrix. Referring to Table 2.1 for the weighting matrices for $Z(p)$ and the ABCD matrix, and substituting into (12), one obtains the well known relation

$$Z(p) = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right\}\left\{ \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}\right\}^{-1} = \begin{bmatrix} \frac{A}{C} & \frac{AD-BC}{C} \\ \frac{1}{C} & \frac{D}{C} \end{bmatrix}. \tag{18}$$

III. GENERALIZED RECIPROCITY CRITERION FOR N-PORT NETWORKS

An n-port network is said to be reciprocal if for any two sets of voltage and current vectors $V'(p)$, $I'(p)$ and $V''(p)$ and $I''(p)$ which are compatible with the network, the reciprocity relation

$$V'^{t}(p)\; I''(p) \equiv V''^{t}(p)\; I'(p) \tag{19}$$

is satisfied.

Observe that in order to apply the above definition for testing the reciprocity of an n-port network, the identity (19) must be satisfied for all possible voltage-current vector pairs compatible with the network. This is clearly a formidable if not impossible task for most networks. The following theorem, whose proof is given in [4], gives a simple criterion for testing the reciprocity of any n-port network characterized by any canonic matrix $T(p)$ and its associated weighting submatrices a, b, c, and d:

Theorem 1 A linear n-port network is reciprocal if and only if the Reciprocity Matrix

$$R(p) \equiv [c\;T(p) + d]^{t}\;[a\;T(p) + b] \tag{20}$$

is a symmetric matrix.

In view of theorem 1, one can readily prove the following corollary [4]:

Corollary to Theorem 1

An n-port network characterized by an open-circuit impedance matrix $Z(p)$, short-circuit admittance matrix $Y(p)$, scattering matrix $\bar{S}(p)$ (normalized to complex port numbers) or scattering matrix $S(p)$ (normalized to real port numbers) is reciprocal if and only if $Z(p)$, $Y(p)$ $\bar{S}(p)$ or $S(p)$ is a symmetric matrix.

The above corollary seems to suggest that the canonic matrix $T(p)$ of any reciprocal n-port must be a symmetric matrix. This, however, is not true, for consider again the ABCD 2-port parameters with $T(p) = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$.

Substituting the weighting submatrices from Table 2.1 into (20), one obtains the reciprocity matrix

$$\underset{\sim}{R}(p) = \left\{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}\right\}^t \left\{\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right\} = \begin{bmatrix} CA & CB \\ DA-1 & DB \end{bmatrix} . \tag{21}$$

Hence by Theorem 1, the 2-port characterized by the ABCD canonic matrix is reciprocal if and only if CB = DA-1, or AD - BC = 1, which is the familiar condition for reciprocity. This example clearly illustrates the fact that the canonic matrix $\underset{\sim}{T}(p)$ of a reciprocal n-port network need not be a symmetric matrix; it is the reciprocity matrix which has to be symmetric.

As an illustration of the utility of Theorem 1, consider the following examples:

Example 1 Ideal Transformer From table 3, one obtains

$$\underset{\sim}{R} = \left\{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} n & 0 \\ 0 & -1/n \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right\}^t \left\{\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} n & 0 \\ 0 & -1/n \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right\} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} , \tag{22}$$

hence, an ideal transformer is a reciprocal 2-port.

Example 2 Nullor From Table 2.2, one obtains

$$\underset{\sim}{R} = \left\{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix}\right\}^t \left\{\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix}\right\} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} , \tag{23}$$

hence, the nullor is a non-reciprocal 2-port

IV. GENERALIZED PASSIVITY CRITERION FOR N-PORT NETWORKS

Substituting $\underset{\sim}{V}(j\omega)$ and $\underset{\sim}{I}(j\omega)$ from (5) into the expression for total complex power flowing into an n-port network, i.e.,

$$P_T = \frac{1}{2} \underset{\sim}{V}^{\dagger}(j\omega)\, \underset{\sim}{I}(j\omega) \quad , \tag{24}$$

one obtains the relation

$$P_T = \frac{1}{2} \underset{\sim}{X}_b^{\dagger}(j\omega)\, \underset{\sim}{D}(j\omega)\, \underset{\sim}{X}_b(j\omega) \tag{25}$$

where $\underset{\sim}{D}(j\omega) \equiv [\underset{\sim}{c}\, \underset{\sim}{T}(j\omega) + \underset{\sim}{d}]^{\dagger} [\underset{\sim}{a}\, \underset{\sim}{T}(j\omega) + \underset{\sim}{b}]$ (26)

The matrix $\underset{\sim}{D}(j\omega)$ completely characterizes the passivity and losslessness of an n-port network and is called the Characteristic Matrix. Observe that for frequency invariant n-port networks, the characteristic matrix and the reciprocity matrix are identical. From (25), one readily obtains the expression for the average real power dissipated in any n-port network as

$$P_{av} = \frac{1}{4} \underset{\sim}{X}_b^{\dagger}(j\omega)\, \underset{\sim}{D}_H(j\omega)\, \underset{\sim}{X}_b(j\omega) \tag{27}$$

where $\underset{\sim}{D}_H(j\omega) \equiv 1/2\ [\underset{\sim}{D}(j\omega) + \underset{\sim}{D}^{\dagger}(j\omega)]$ (28)

is the Hermitian part of the characteristic matrix $\underset{\sim}{D}(j\omega)$.

In view of (27), one can now state the following two theorems regarding the losslessness and passivity of an n-port network:

Theorem 2

The necessary and sufficient condition that a linear n-port network be lossless is that its characteristic matrix $D(j\omega)$ be a skew-Hermitian Matrix for all real frequency ω, i.e., $\underset{\sim}{D}_H(j\omega)$ must be equal identically to the null matrix.

Theorem 3

A necessary condition that a linear n-port network be passive is that the Hermitian part $\underset{\sim}{D}_H(j\omega)$ of its characteristic matrix be a positive-semi-definite matrix for all real frequency ω.

As an application of the above theorems, consider the following examples:

Example 3 Suppose it is desired to find the losslessness criterion in terms of the scattering matrix $\underset{\sim}{S}(p)$, normalized to complex port numbers. Substituting the weighting submatrices from Table 1 into (26) and (28), one obtains

$$\underset{\sim}{D}_H(j\omega) = \underset{\sim}{1}_n - \underset{\sim}{\bar{S}}^{\dagger}(j\omega)\,\underset{\sim}{\bar{S}}(j\omega) \ . \tag{29}$$

Hence in view of theorem 2, one obtains the well known result that an n-port network is lossless if and only if its scattering matrix $\underset{\sim}{\bar{S}}(j\omega)$ is a unitary matrix [5].

Example 4 Consider again the Nullor of Example 2. Since the Nullor is a frequency invariant 2-port, its characteristic matrix $\underset{\sim}{D}$ is identical to the reciprocity matrix $\underset{\sim}{R}$ given by (23). Hence

$$\underset{\sim}{D}_H = \frac{1}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \ . \tag{30}$$

Since det $\underset{\sim}{D}_H = -1/2$, Theorem 3 implies that the nullor is an active 2-port.

Since theorem 3 gives only a necessary condition for passivity, it is not useful as a passivity criterion. The real test for passivity is given by the following theorem whose lengthy proof is given in [4].

Theorem 4 Let the characteristic matrix $\underset{\sim}{D}(p)$ (or its transpose conjugate $\underset{\sim}{D}^{\dagger}(p)$) of an n-port network $\mathfrak{N}$ be analytic in the right half p-plane. Then $\mathfrak{N}$ is passive if and only if $\underset{\sim}{D}(p)$ (or $\underset{\sim}{D}^{\dagger}(p)$) is a positive-real matrix.

The above theorem is very general and among other things, it asserts that an n-port network $\mathfrak{N}$ is passive if and only if its open-circuit impedance matrix or short-circuit admittance matrix is a positive real matrix.

V. INTERCONNECTION OF N-PORT NETWORKS

It is often desired to characterize a composite n-port network made up of an interconnection of two or more multi-port (not necessarily n) networks. Since the present characterization does not favor any particular weighting matrix, it is especially suitable to handle this type of problems. To illustrate the flexibility of the present approach, suppose it is desired to find the open-circuit impedance matrix of the composite network consisting of an interconnection of two n-port networks $\mathfrak{N}'$ and $\mathfrak{N}''$, as shown in Fig. 1. It is shown in [4] that a convenient choice of weighting matrices $\underset{\sim}{W}'$ and $\underset{\sim}{W}''$ for this network are:

$$\underset{\sim}{W}' = \left[\begin{array}{cccc:cccc} 0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&1&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1 \\ \hdashline 1&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&1&0&0&0&0 \end{array}\right] \quad \text{and} \quad \underset{\sim}{W}'' = \left[\begin{array}{cccc:cccc} 0&0&0&0&1&0&0&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&1 \\ \hdashline 1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0 \end{array}\right]$$

If $\underset{\sim}{T}'(P)$ and $\underset{\sim}{T}''(p)$ are the canonic matrices corresponding to the above weighting matrices $\underset{\sim}{W}'$ and $\underset{\sim}{W}''$; if t_{jk}' and t_{jk}'' denote the jkth element of $\underset{\sim}{T}'(p)$ and $\underset{\sim}{T}''(p)$, respectively, then the open-circuit impedance matrix of the composite network is readily found to be

$$\underset{\sim}{Z}(p) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ (t_{31}'-t_{21}'') & -(t_{33}'+t_{22}'') & z_{33} & z_{34} \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} (t_{11}'+t_{11}'') & (t_{12}''-t_{13}') & (t_{11}'+t_{13}'') & z_{14}' \\ (t_{21}'-t_{41}'') & -(t_{23}'+t_{42}'') & z_{23}' & z_{24}' \\ 0 & -1 & 0 & 0 \\ z_{41}' & z_{42}' & z_{43}' & z_{44}' \end{bmatrix}^{-1}$$

where $z_{33} = (t_{31}'+t_{34}'-t_{23}'')$; $z_{34} = -(t_{32}'+t_{24}'')$; $z_{14}' = (t_{14}''-t_{12}')$; $z_{23}' = (t_{21}'+t_{24}'-t_{43}'')$; $z_{24}' = (-t_{22}'-t_{44}')$; $z_{41}' = (t_{41}'-t_{11}''+t_{31}'')$; $z_{42}' = (-t_{43}'-t_{12}''+t_{32}'')$; $z_{43}' = (t_{41}'+t_{44}'-t_{13}''+t_{33}'')$ and $z_{44} = (-t_{14}'-t_{14}''+t_{34}'')$.

VI. CASCADING OF N-PORT NETWORKS

Closely related to the interconnection problem is the cascading of two or more n-port networks. In order to simplify analysis, it is often desirable to find a suitable characterization so that the matrix of the composite n-port network is simply the product of the individual matrices. For example, the ABCD matrix for 2-port networks and the transfer scattering matrix as specified in Table 2.2 both possess the above property. It is therefore important to ask the question: "Are there any other matrices possessing the above property and, if so, what are the general forms of these matrices?" The answer to this question is given in the following theorem [4]:

Theorem 5 Let $\mathfrak{N}$ be an n-port network resulting from the cascade connection of two n-port networks $\mathfrak{N}'$ and $\mathfrak{N}''$ as shown in Fig. 2. Let $\mathfrak{N}$, $\mathfrak{N}'$ and $\mathfrak{N}''$ be characterized respectively by $\underset{\sim}{T}(p)$, $\underset{\sim}{W}$; $\underset{\sim}{T}'(p)$, $\underset{\sim}{W}'$; and $\underset{\sim}{T}''(p)$, $\underset{\sim}{W}''$. Then a sufficient condition that $\underset{\sim}{T}(p) = \underset{\sim}{T}'(p)\,\underset{\sim}{T}''(p)$ is that the inverse weighting matrices be of the following forms:

$$\underset{\sim}{W}'^{-1} = \left[\begin{array}{cc|cc} \underset{\sim}{\alpha}'_{11} & \underset{\sim}{0} & \underset{\sim}{\beta}'_{11} & \underset{\sim}{0} \\ \underset{\sim}{\alpha}'_{21} & \underset{\sim}{0} & \underset{\sim}{\beta}'_{21} & \underset{\sim}{0} \\ \hline \underset{\sim}{0} & \underset{\sim}{\mu}_1 & \underset{\sim}{0} & \underset{\sim}{\mu}_3 \\ \underset{\sim}{0} & \underset{\sim}{\mu}_2 & \underset{\sim}{0} & \underset{\sim}{\mu}_4 \end{array}\right], \quad \underset{\sim}{W}''^{-1} = \left[\begin{array}{cc|cc} \underset{\sim}{\mu}_1 & \underset{\sim}{0} & -\underset{\sim}{\mu}_3 & \underset{\sim}{0} \\ \underset{\sim}{\mu}_2 & \underset{\sim}{0} & -\underset{\sim}{\mu}_4 & \underset{\sim}{0} \\ \hline \underset{\sim}{0} & \underset{\sim}{\gamma}''_{12} & \underset{\sim}{0} & \underset{\sim}{\delta}''_{12} \\ \underset{\sim}{0} & \underset{\sim}{\gamma}''_{22} & \underset{\sim}{0} & \underset{\sim}{\delta}''_{22} \end{array}\right] \text{ and } \underset{\sim}{W}^{-1} = \left[\begin{array}{cc|cc} \underset{\sim}{\alpha}'_{11} & \underset{\sim}{0} & \underset{\sim}{\beta}'_{11} & \underset{\sim}{0} \\ \underset{\sim}{\alpha}'_{21} & \underset{\sim}{0} & \underset{\sim}{\beta}'_{21} & \underset{\sim}{0} \\ \hline \underset{\sim}{0} & \underset{\sim}{\gamma}''_{12} & \underset{\sim}{0} & \underset{\sim}{\delta}''_{12} \\ \underset{\sim}{0} & \underset{\sim}{\gamma}''_{22} & \underset{\sim}{0} & \underset{\sim}{\delta}''_{22} \end{array}\right] \qquad (32)$$

Observe that in order for $\underset{\sim}{T}(p) = \underset{\sim}{T}'(p)\underset{\sim}{T}''(p)$, the weighting matrices $\underset{\sim}{W}$, $\underset{\sim}{W}'$ and $\underset{\sim}{W}''$ need not be identical. Only certain strategically placed elements need be chosen to satisfy the above conditions. In the special case where $\underset{\sim}{W} = \underset{\sim}{W}' = \underset{\sim}{W}''$, one can state the following corollary:

Corollary to Theorem 5 If $\underset{\sim}{W} = \underset{\sim}{W}' = \underset{\sim}{W}''$, then the condition that $\underset{\sim}{T}(p) = \underset{\sim}{T}'(p)\,\underset{\sim}{T}''(p)$ is

$$\underset{\sim}{W}^{-1} = \underset{\sim}{W}'^{-1} = \underset{\sim}{W}''^{-1} = \left[\begin{array}{cc|cc} \underset{\sim}{\mu}_1 & \underset{\sim}{0} & -\underset{\sim}{\mu}_3 & \underset{\sim}{0} \\ \underset{\sim}{\mu}_2 & \underset{\sim}{0} & -\underset{\sim}{\mu}_4 & \underset{\sim}{0} \\ \hline \underset{\sim}{0} & \underset{\sim}{\mu}_1 & \underset{\sim}{0} & \underset{\sim}{\mu}_3 \\ \underset{\sim}{0} & \underset{\sim}{\mu}_2 & \underset{\sim}{0} & \underset{\sim}{\mu}_4 \end{array}\right] . \qquad (33)$$

It is easily verified that the transfer scattering matrix satisfies (32) and the ABCD matrix satisfies (33). Another set of weighting matrices which have never been used but which satisfies (33) is given by

$$\underset{\sim}{W}^{-1} = \underset{\sim}{W}'^{-1} = \underset{\sim}{W}''^{-1} = \left[\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ \hline 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 1 \end{array}\right] . \qquad (34)$$

VII. FORMAL SYNTHESIS OF LINEAR N-PORT NETWORKS

A great deal of work has been done in the formal synthesis of linear n-port networks [6-10]. The basic philosophy of these works usually consists of first transforming the given specification into an equivalent open-circuit impedance matrix $\underset{\sim}{Z}(p)$ or short-circuit admittance matrix $\underset{\sim}{Y}(p)$, and then realize an n-port with the prescribed $\underset{\sim}{Z}(p)$ or $\underset{\sim}{Y}(p)$. However, the above procedure would work only if $\underset{\sim}{Z}(p)$ or $\underset{\sim}{Y}(p)$ exists. Hence from a network theoretic point of view, it is desirable to find a universal synthesis method, so that specification can be realized directly without having to convert from one specification into another. The general method for synthesis is best described by the following theorem:

Theorem 6 A Universal Synthesis Method **Any n-port network characterized by an nxn canonic matrix $\underset{\sim}{T}(p)$ and its associated weighting matrix $\underset{\sim}{W}$ is realizable by the canonic n-port model shown in Fig. 3.**

The controlled sources with controlling coefficients which are functions of p may be realized by an equivalent network containing controlled sources with real controlling coefficients, resistors and capacitors. In particular, the 2n-port network $\mathfrak{N}'$, shown in Fig. 3 may be realized by a network with voltage transfer matrix $\underset{\sim}{T}(p)$ using techniques such as those given in [11-12] or by the network shown in Fig. 4, using any of the techniques given in [6-11].

The proof of theorem 6 follows directly from the definition of the canonic and weighting matrices. Observe that the network given in Fig. 3 is a canonic model in the sense that networks of any characterization can be realized by this model. As an illustration of the utility of this model, the nullor is synthesized in three steps, as shown in Fig.5.

REFERENCES

1. V. Belevitch, "Four Dimensional Transformations of 4-pole Matrices with Applications to the Synthesis of Reactance 4-Poles", IRE Trans. On Circuit Theory, June, 1956, pp. 105-111.
2. H. J. Carlin, "Singular Network Elements", IEEE Trans. on Circuit Theory, March, 1964, pp. 67-72.
3. H. Goldstein, Classical Mechanics, Addison-Wesly Publishing Co., Inc., Reading, Mass, 1959.
4. L. O. Chua and B. J. Leon, "A Unified N-port System Theory", Purdue University Report TR-EE65-18, September, 1965.
5. R. A. Rohrer, "The Scattering Matrix: Normalized to Complex n-port Load Networks", IEEE Trans. On Circuit Theory, June, 1965, pp. 223-229.
6. B. McMillan, "Introduction To Formal Realizability Theory", Bell System Tech. Journal, vol. 31, March and May, 1952, pp. 217-279 and 541-600.
7. Y. Oono and K. Yasuura, "Synthesis of Finite Passive 2n-terminal Networks with Prescribed Scattering Matrices", Mem. Kyushu Univ. (Faculty of Engineering), Vol. 14, no. 2, pp. 125-177, 1954.
8. W. Sandberg, "Synthesis of N-Port Active RC Networks," Bell System Tech. Journal, Aug. 16, 1960, pp. 329-347.
9. R. J. Duffin, D. Hazony N. Morrison, "Network Synthesis Through Hybrid Matrices", Case Institute of Technology Report No. 4, ARCRL 63-568, Aug. 18, 1964.
10. D. C. Youla, "Cascade Synthesis of Passive n-ports", Report RADC-TDR-64-332, Rome Air Development Center, Aug. 1964.
11. P. M. Lewis, "The Concept of the One in Voltage Transfer Synthesis," IRE Trans. On Circuit Theory, December, 1955, pp. 316-319.
12. A. Fialkow, D. Hazony, P. Kodali and T. Zeren, "Transfer Function Realizability of Multiport RLC Transformerless Grounded Networks," IEEE Trans. On Circuit Theory, March, 1964, pp. 73-80.

	W				W^{-1}			
	a	b	c	d	α	β	γ	δ
$Z(p)$	1_n	0_n	0_n	1_n	$\alpha = a$	$\beta = b$	$\gamma = c$	$\delta = d$
$Y(p)$	0_n	1_n	1_n	0_n	$\alpha = a$	$\beta = b$	$\gamma = c$	$\delta = d$
$H(p)$	$1_k \dotplus 0_{n-k}$	$0_k \dotplus 1_{n-k}$	$0_k \dotplus 1_{n-k}$	$1_k \dotplus 0_{n-k}$	$\alpha = a$	$\beta = b$	$\gamma = c$	$\delta = d$
$\bar{S}(p)$	$r^{\frac{1}{2}} + \frac{1}{2} r^{-\frac{1}{2}}(z - z^*)$	$r^{\frac{1}{2}} - \frac{1}{2} r^{-\frac{1}{2}}(z - z^*)$	$-r^{-\frac{1}{2}}$	$r^{-\frac{1}{2}}$	$\frac{1}{2} r^{-\frac{1}{2}}$	$-\frac{1}{2} z^* r^{-\frac{1}{2}}$	$\frac{1}{2} r^{-\frac{1}{2}}$	$\frac{1}{2} z r^{-\frac{1}{2}}$
$S(p)$	$r^{\frac{1}{2}}$	$r^{\frac{1}{2}}$	$-r^{-\frac{1}{2}}$	$r^{-\frac{1}{2}}$	$\frac{1}{2} r^{-\frac{1}{2}}$	$-\frac{1}{2} r^{\frac{1}{2}}$	$\frac{1}{2} r^{-\frac{1}{2}}$	$\frac{1}{2} r^{\frac{1}{2}}$

$Z(p)$ = open-circuit Impedence Matrix, i.e. $V = Z I$

$Y(p)$ = Short-circuit Admittance Matrix, i.e. $I = Y V$

$H(p)$ = Hybrid Matrix $\begin{bmatrix} V_k \\ I_{n-k} \end{bmatrix} = \begin{bmatrix} H \end{bmatrix} \begin{bmatrix} I_k \\ V_{n-k} \end{bmatrix}$

r = diag $(r_1, r_2, \cdots r_n)$

$r^{\pm\frac{1}{2}}$ = diag $(r_1^{\pm\frac{1}{2}}, r_2^{\pm\frac{1}{2}}, \cdots r_n^{\pm\frac{1}{2}})$

z = diag $(z_1, z_2, \cdots z_n)$

z^* = diag $(z_1^*, z_2^*, \cdots z_n^*)$

$\bar{S}(p)$ = Voltage Basis Scattering Matrix Normalized to Complex Port Numbes $z_1, z_2 \cdots z_n$, Where $z_i = r_i + jx_i$

$S(p)$ = Voltage Basis Scattering Matrix Normalized to Real Numbers $r_1, r_2, \cdots r_n$

0_n = n x n Null Matrix

$\dotplus$ denotes direct sum

1_n = n x n Unit Matrix

TABLE I

WEIGHTING AND INVERSE WEIGHTING MATRICES OF SOME COMMON n-PORT NETWORK CHARACTERIZATIONS

	In Terms of W and $I(p)$	In Terms of W^{-1} and $I(p)$
	Let $N(p) \equiv [a I(p) + b][c I(p) + d]^{-1}$	Let $M(p) \equiv [\alpha - I(p)\gamma]^{-1}[I(p)\delta - \beta]$
$Z(p)$	$Z(p) = N(p)$	$Z(p) = M(p)$
$Y(p)$	$Y(p) = N^{-1}(p)$	$Y(p) = M^{-1}(p)$
$\bar{S}(p)$	$\bar{S}(p) = \{1_n + r^{\frac{1}{2}} N^{-1}(p) r^{\frac{1}{2}}[1_n + \frac{1}{2}(z - z^*) r^{-1}]]\}^{-1}$ $\{1_n - r^{\frac{1}{2}} N^{-1}(p) r^{\frac{1}{2}}[1_n - \frac{1}{2}(z - z^*) r^{-1}]\}$	$\bar{S}(p) = \{1_n - z^* r^{-\frac{1}{2}} M^{-1}(p) r^{\frac{1}{2}}\}\{1_n + z r^{-\frac{1}{2}} M^{-1}(p) r^{\frac{1}{2}}\}^{-1}$
	$\bar{S}(p) = \{r^{-\frac{1}{2}} N(p) r^{-\frac{1}{2}} + \frac{1}{2}(z - z^*) r^{-1} + 1_n\}^{-1}\{r^{-\frac{1}{2}} N(p) r^{-\frac{1}{2}} + \frac{1}{2}(z - z^*) r^{-1} - 1_n\}$	$\bar{S}(p) = \{r^{-\frac{1}{2}} M(p) r^{-\frac{1}{2}} - z^* r^{-1}\}\{r^{-\frac{1}{2}} M(p) r^{-\frac{1}{2}} + z r^{-1}\}^{-1}$
$S(p)$	$S(p) = \{r^{-\frac{1}{2}} N(p) r^{-\frac{1}{2}} + 1_n\}^{-1}\{r^{-\frac{1}{2}} N(p) r^{-\frac{1}{2}} - 1_n\}$	$S(p) = \{r^{-\frac{1}{2}} M(p) r^{-\frac{1}{2}} - 1_n\}\{r^{-\frac{1}{2}} M(p) r^{-\frac{1}{2}} + 1_n\}^{-1}$
	$S(p) = \{1_n + r^{\frac{1}{2}} N^{-1}(p) r^{\frac{1}{2}}\}^{-1}\{1_n - r^{\frac{1}{2}} N^{-1}(p) r^{\frac{1}{2}}\}$	$S(p) = \{1_n - r^{\frac{1}{2}} M^{-1}(p) r^{\frac{1}{2}}\}\{1_n + r^{\frac{1}{2}} M^{-1}(p) r^{\frac{1}{2}}\}^{-1}$

TABLE 4.1

COMMON n-PORT NETWORK CHARACTERIZATIONS IN TERMS OF THE CANONIC MATRIX

		$I(p),\ a,\ b,\ c,\ d$	$I(p),\ \alpha,\ \beta,\ \gamma,\ \delta$
In Terms of	$Z(p)$	$I(p) = [Z(p)c - a]^{-1}[b - Z(p)d]$	$I(p) = [\alpha Z(p) + \beta][\gamma Z(p) + \delta]^{-1}$
	$Y(p)$	$I(p) = [Y(p)a - c]^{-1}[d - Y(p)b]$	$I(p) = [\alpha + \beta Y(p)][\gamma + \delta Y(p)]^{-1}$
	$\bar{S}(p)$	$I(p) = \{[1_n - \bar{S}(p)] r^{-\frac{1}{2}} a - [z^* + \bar{S}(p) z] r^{-\frac{1}{2}} c\}^{-1}$ $\{[\bar{S}(p) - 1_n] r^{-\frac{1}{2}} b + [z^* + \bar{S}(p) z] r^{-\frac{1}{2}} d\}$	$I(p) = \{\alpha r^{\frac{1}{2}}[1_n - \bar{S}(p)]^{-1}[z^* + \bar{S}(p) z] r^{-\frac{1}{2}} + \beta\}$ $\{\delta + \gamma r^{\frac{1}{2}}[1_n - \bar{S}(p)]^{-1}[z^* + \bar{S}(p) z] r^{-\frac{1}{2}}\}^{-1}$
	$\bar{S}(p)$	$I(p) = \{r^{-\frac{1}{2}}[(1_n + \bar{S}(p))(1_n - \bar{S}(p))^{-1} - \frac{1}{2}(z - z^*) r^{-1}]^{-1} r^{-\frac{1}{2}} a - c\}^{-1}$ $\{d - r^{-\frac{1}{2}}[(1_n + \bar{S}(p))(1_n - \bar{S}(p))^{-1} - \frac{1}{2}(z - z^*) r^{-1}]^{-1} r^{\frac{1}{2}} b\}$	$I(p) = \{\alpha + \beta r^{\frac{1}{2}}[\bar{S}(p) z + z^*]^{-1}[1_n - \bar{S}(p)] r^{-\frac{1}{2}}\}$ $\{\gamma + \delta r^{\frac{1}{2}}[\bar{S}(p) z + z^*]^{-1}[1_n - \bar{S}(p)] r^{-\frac{1}{2}}\}^{-1}$
	$\bar{S}(p)$	$I(p) = \{r^{\frac{1}{2}}[(\bar{S}(p) + 1_n)(1_n - \bar{S}(p))^{-1} - \frac{1}{2} r^{-1}(z - z^*)] r^{\frac{1}{2}} c - a\}^{-1}$ $\{b - r^{\frac{1}{2}}[(\bar{S}(p) + 1_n)(1_n - \bar{S}(p))^{-1} - \frac{1}{2} r^{-1}(z - z^*)] r^{\frac{1}{2}} d\}$	
	$S(p)$	$I(p) = \{[1_n - S(p)] r^{-\frac{1}{2}} a - [1_n + S(p)] r^{\frac{1}{2}} c\}^{-1}$ $\{[S(p) - 1_n] r^{-\frac{1}{2}} b + [S(p) + 1_n] r^{\frac{1}{2}} d\}$	$I(p) = \{\alpha r^{\frac{1}{2}}[1_n - S(p)]^{-1}[1_n + S(p)] r^{\frac{1}{2}} + \beta\}$ $\{\delta + \gamma r^{\frac{1}{2}}[1_n - S(p)]^{-1}[1_n + S(p)] r^{\frac{1}{2}}\}$
	$S(p)$	$I(p) = \{r^{-\frac{1}{2}}[1_n + S(p)][1_n - S(p)]^{-1} r^{\frac{1}{2}} a - c\}^{-1}$ $\{d - r^{-\frac{1}{2}}[1_n + S(p)][1_n - S(p)]^{-1} r^{\frac{1}{2}} b\}$	$I(p) = \{\alpha + \beta r^{-\frac{1}{2}}[S(p) + 1_n]^{-1}[1_n - S(p)] r^{\frac{1}{2}}\}$ $\{\delta r^{-\frac{1}{2}}[S(p) + 1_n]^{-1}[1_n - S(p)] r^{\frac{1}{2}} + \gamma\}^{-1}$
	$S(p)$	$I(p) = \{r^{\frac{1}{2}}[S(p) + 1_n][1_n - S(p)]^{-1} r^{\frac{1}{2}} c - a\}^{-1}$ $\{b - r^{\frac{1}{2}}[S(p) + 1_n][1_n - S(p)]^{-1} r^{-\frac{1}{2}} d\}$	

TABLE 4.2

CANONIC MATRIX IN TERMS OF SOME COMMON n-PORT NETWORK CHARACTERIZATIONS

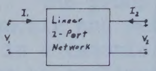
	$\underset{\sim}{W} = \underset{\sim}{W}^{-1}$			
	$\underset{\sim}{a} = \underset{\sim}{\alpha}$	$\underset{\sim}{b} = \underset{\sim}{\beta}$	$\underset{\sim}{c} = \underset{\sim}{\gamma}$	$\underset{\sim}{d} = \underset{\sim}{\delta}$
$\begin{bmatrix}V_1\\V_2\end{bmatrix} = \begin{bmatrix}z_{11}&z_{12}\\z_{21}&z_{22}\end{bmatrix}\begin{bmatrix}I_1\\I_2\end{bmatrix}, \underset{\sim}{T} = \underset{\sim}{Z} = \begin{bmatrix}z_{11}&z_{12}\\z_{21}&z_{22}\end{bmatrix}$ Open-circuit Parameters	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$
$\begin{bmatrix}I_1\\I_2\end{bmatrix} = \begin{bmatrix}Y_{11}&Y_{12}\\Y_{21}&Y_{22}\end{bmatrix}\begin{bmatrix}V_1\\V_2\end{bmatrix}, \underset{\sim}{T} = \underset{\sim}{Y} = \begin{bmatrix}Y_{11}&Y_{12}\\Y_{21}&Y_{22}\end{bmatrix}$ Short-circuit Parameters	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$
$\begin{bmatrix}V_1\\I_1\end{bmatrix} = \begin{bmatrix}A&B\\C&D\end{bmatrix}\begin{bmatrix}V_2\\-I_2\end{bmatrix}, \underset{\sim}{T} = \begin{bmatrix}A&B\\C&D\end{bmatrix}$ ABCD Transmission Parameters	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&-1\end{bmatrix}$
$\begin{bmatrix}I_2\\V_2\end{bmatrix} = \begin{bmatrix}A&C\\B&D\end{bmatrix}\begin{bmatrix}-I_1\\V_1\end{bmatrix}, \underset{\sim}{T} = \begin{bmatrix}A&C\\B&D\end{bmatrix}$ Inverse Transmission Parameters	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\begin{bmatrix}-1&0\\0&0\end{bmatrix}$
$\begin{bmatrix}V_1\\I_2\end{bmatrix} = \begin{bmatrix}h_{11}&h_{12}\\h_{21}&h_{22}\end{bmatrix}\begin{bmatrix}I_1\\V_2\end{bmatrix}, \underset{\sim}{T} = \underset{\sim}{H} = \begin{bmatrix}h_{11}&h_{12}\\h_{21}&h_{22}\end{bmatrix}$ Hybrid Parameters	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$
$\begin{bmatrix}I_1\\V_2\end{bmatrix} = \begin{bmatrix}g_{11}&g_{12}\\g_{21}&g_{22}\end{bmatrix}\begin{bmatrix}V_1\\I_2\end{bmatrix}, \underset{\sim}{T} = \begin{bmatrix}g_{11}&g_{12}\\g_{21}&g_{22}\end{bmatrix}$ g - Parameters	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$

TABLE 2.1

WEIGHTING AND INVERSE WEIGHTING MATRICES OF COMMON 2-PORT PARAMETERS (CASE I: $\underset{\sim}{W} = \underset{\sim}{W}^{-1}$)

	Scattering Matrix Normalized to z_1, z_2	Scattering Matrix Normalized to r_1, r_2	Transfer Scattering Matrix	Nullor
	$\underset{\sim}{T} = \underset{\sim}{\bar{S}} = \begin{bmatrix}\bar{s}_{11}&\bar{s}_{12}\\\bar{s}_{21}&\bar{s}_{22}\end{bmatrix}$	$\underset{\sim}{T} = \underset{\sim}{S} = \begin{bmatrix}s_{11}&s_{12}\\s_{21}&s_{22}\end{bmatrix}$	$\underset{\sim}{T} = \underset{\sim}{\mathcal{T}} = \begin{bmatrix}\tau_{11}&\tau_{12}\\\tau_{21}&\tau_{22}\end{bmatrix}$	$\underset{\sim}{T} = \underset{\sim}{\mathcal{A}} = \begin{bmatrix}1&0\\0&1\end{bmatrix}$
$\underset{\sim}{a}$	$\begin{bmatrix}\frac{r_1+jx_1}{\sqrt{r_1}}&0\\0&\frac{r_2+jx_2}{\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}\sqrt{r_1}&0\\0&\sqrt{r_2}\end{bmatrix}$	$\begin{bmatrix}\sqrt{r_1}&\sqrt{r_1}\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$
$\underset{\sim}{b}$	$\begin{bmatrix}\frac{r_1-jx_1}{\sqrt{r_1}}&0\\0&\frac{r_2-jx_2}{\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}\sqrt{r_1}&0\\0&\sqrt{r_2}\end{bmatrix}$	$\begin{bmatrix}0&0\\\sqrt{r_2}&\sqrt{r_2}\end{bmatrix}$	$\begin{bmatrix}-1&0\\1&0\end{bmatrix}$
$\underset{\sim}{c}$	$\begin{bmatrix}-\frac{1}{\sqrt{r_1}}&0\\0&-\frac{1}{\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}-\frac{1}{\sqrt{r_1}}&0\\0&-\frac{1}{\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}-\sqrt{r_1}&\sqrt{r_1}\\0&0\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$
$\underset{\sim}{d}$	$\begin{bmatrix}\frac{1}{\sqrt{r_1}}&0\\0&\frac{1}{\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}\frac{1}{\sqrt{r_1}}&0\\0&\frac{1}{\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}0&0\\\sqrt{r_2}&-\sqrt{r_2}\end{bmatrix}$	$\begin{bmatrix}0&-1\\0&1\end{bmatrix}$
$\underset{\sim}{\alpha}$	$\begin{bmatrix}\frac{1}{2\sqrt{r_1}}&0\\0&\frac{1}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}\frac{1}{2\sqrt{r_1}}&0\\0&\frac{1}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}\frac{1}{2\sqrt{r_1}}&0\\\frac{1}{2\sqrt{r_1}}&0\end{bmatrix}$	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$
$\underset{\sim}{\beta}$	$\begin{bmatrix}\frac{-(r_1-jx_1)}{2\sqrt{r_1}}&0\\0&\frac{-(r_2-jx_2)}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}-\frac{1}{2}\sqrt{r_1}&0\\0&-\frac{1}{2}\sqrt{r_2}\end{bmatrix}$	$\begin{bmatrix}-\frac{1}{2\sqrt{r_1}}&0\\\frac{1}{2\sqrt{r_1}}&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&1\end{bmatrix}$
$\underset{\sim}{\gamma}$	$\begin{bmatrix}\frac{1}{2\sqrt{r_1}}&0\\0&\frac{1}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}\frac{1}{2\sqrt{r_1}}&0\\0&\frac{1}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}0&\frac{1}{2\sqrt{r_2}}\\0&\frac{1}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$
$\underset{\sim}{\delta}$	$\begin{bmatrix}\frac{r_1+jx_1}{2\sqrt{r_1}}&0\\0&\frac{r_2+jx_2}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}\frac{1}{2}\sqrt{r_1}&0\\0&\frac{1}{2}\sqrt{r_2}\end{bmatrix}$	$\begin{bmatrix}0&\frac{1}{2\sqrt{r_2}}\\0&-\frac{1}{2\sqrt{r_2}}\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$

TABLE 2.2

WEIGHTING AND INVERSE WEIGHTING MATRICES OF COMMON 2-PORT PARAMETERS (CASE 2: $\underset{\sim}{W} \neq \underset{\sim}{W}^{-1}$)

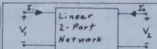
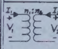
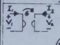
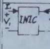
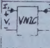
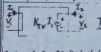
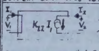
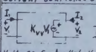
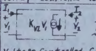
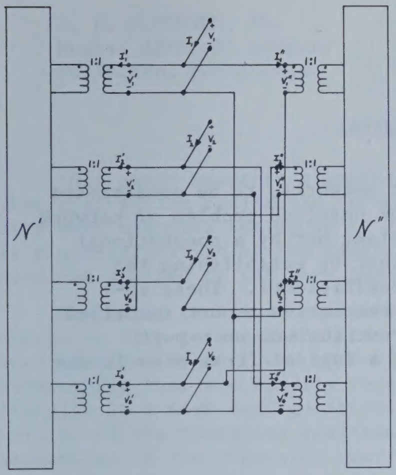
	$\underset{\sim}{W} = \underset{\sim}{W}^{-1}$			
	$\underset{\sim}{a} = \underset{\sim}{\alpha}$	$\underset{\sim}{b} = \underset{\sim}{\beta}$	$\underset{\sim}{c} = \underset{\sim}{\gamma}$	$\underset{\sim}{d} = \underset{\sim}{\delta}$
 $\underset{\sim}{T} = \begin{bmatrix}n_1/n_2&0\\0&-n_2/n_1\end{bmatrix}$ Ideal Transformer	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$
 $\underset{\sim}{T} = \begin{bmatrix}0&G\\-G&0\end{bmatrix} = \underset{\sim}{Y}$ Gyrator	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$
 $\underset{\sim}{T} = \begin{bmatrix}1&0\\0&1/K\end{bmatrix}$ Current-type Negative Impedance Converter	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$
 $\underset{\sim}{T} = \begin{bmatrix}-1/K&0\\0&-1\end{bmatrix}$ Voltage-type Negative Impedance Converter	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$
 $\underset{\sim}{T} = \begin{bmatrix}0&0\\K_{IV}&0\end{bmatrix} = \underset{\sim}{Z}$ Current Controlled Voltage Source	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$
 $\underset{\sim}{T} = \begin{bmatrix}0&0\\K_{II}&0\end{bmatrix}$ Current Controlled Current Source	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$
 $\underset{\sim}{T} = \begin{bmatrix}0&0\\K_{VV}&0\end{bmatrix}$ Voltage Controlled Voltage Source	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$
 $\underset{\sim}{T} = \begin{bmatrix}0&0\\K_{VI}&0\end{bmatrix}$ Voltage Controlled Current Source	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$

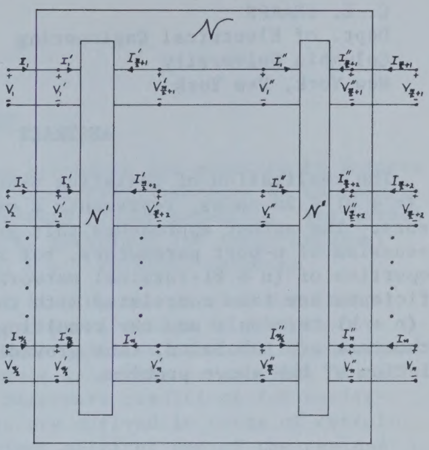
TABLE 3

WEIGHTING AND INVERSE WEIGHTING MATRICES OF COMMON 2-PORT ELEMENTS



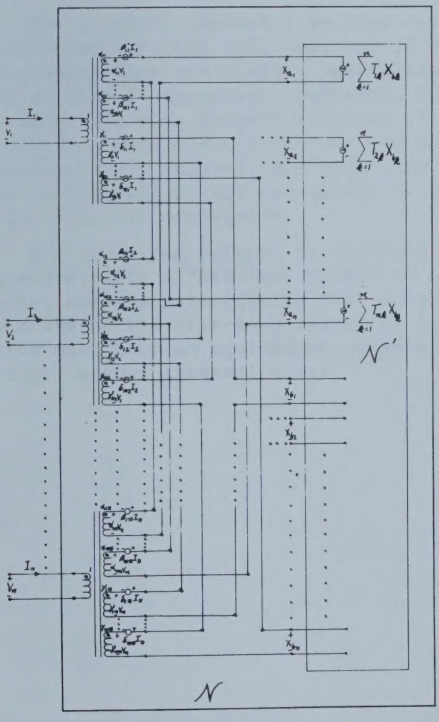
N'
N''
1:1
I_1
V_1
I_1'
V_1'
I_1''
V_1''
I_2
V_2
I_3
V_3
I_n
V_n

FIG. 1 INTERCONNECTION OF TWO n-PORT NETWORKS



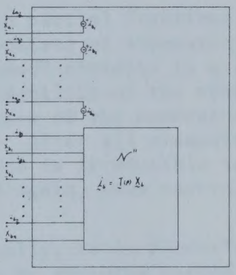
N
N'
N''
I_1
V_1
I_1'
V_1'
I_1''
V_1''
$I_{\frac{n}{2}+1}$
$V_{\frac{n}{2}+1}$
I_2
V_2
$I_{\frac{n}{2}+2}$
$V_{\frac{n}{2}+2}$
$I_{\frac{n}{2}}$
$V_{\frac{n}{2}}$
I_n
V_n

FIG. 2 CASCADE CONNECTION OF TWO n-PORT NETWORKS



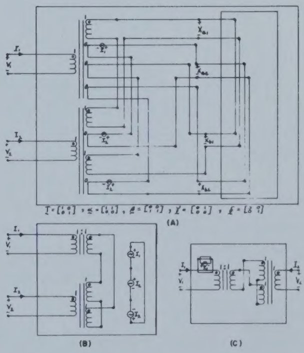
I_1
V_1
I_2
V_2
I_n
V_n
$\beta_{11}I_1$
$\alpha_{11}V_1$
X_{a_1}
$\sum_{k=1}^{n} T_{1k} X_{bk}$
X_{a_2}
$\sum_{k=1}^{n} T_{2k} X_{bk}$
X_{a_n}
$\sum_{k=1}^{n} T_{nk} X_{bk}$
X_{b_1}
X_{b_2}
X_{b_n}
N'
N

FIG. 3 CANONIC MODEL FOR THE SYNTHESIS OF ARBITRARY n-PORT NETWORKS



N''
$\underline{i}_b = \underline{T}(p)\,\underline{X}_b$

FIG. 4 A METHOD FOR REALIZING THE 2n-PORT NETWORK N'



I_1
V_1
I_2
V_2
X_{a_1}
X_{a_2}
X_{b_1}
X_{b_2}
(A)
(B)
(C)
1:1

FIG. 5 REALIZATION OF A NULLOR

RESTRICTIONS ON THE PARAMETERS OF RESISTIVE N-PORT NETWORKS

G. E. SHARPE
Dept. of Electrical Engineering
Columbia University
New York, New York

ABSTRACT

The realization of resistive n-port networks, to be constructed on $(n + k) \leq 2n$ nodes, represents a major unsolved problem of network theory. The author approaches this problem, not by a conventional discussion of n-port parameters, but rather by establishing the properties of $(n + k)$-terminal network coefficients. These coefficients are then correlated with the respective n-port inscribed on $(n + k)$ terminals and the resulting restrictions on n-port parameters are tabulated, thus providing a logical first step in the solution of the above problem.

SYNTHESIS OF GROUNDED RC N-PORTS BY CANONICAL STAR-MESH NETWORKS

J. E. OLIVARES, JR.
Hughes Aircraft Company
Fullerton, California

ABSTRACT

This paper presents a synthesis procedure for grounded RC N-ports that synthesizes realizable terminal immittance matrices as canonical star-mesh networks analogous to alternate Foster networks. The method is a generalization of the Ozaki-Lucal technique for three-terminal networks applied to the more general N + 1 terminal network.

The basic technique consists of finding the partial fraction expansion of a terminal admittance and partitioning it into two or more admittance matrices. One of these matrices consists of elemental subnetworks that are readily realizable as series or parallel "tuned" circuits in a mesh configuration. Necessary conditions for realizability of the remaining admittances are derived in terms of certain properties of the numerator coefficient matrices and of the residue matrices. After all elemental subnetworks are removed, the remaining admittances are irreducible in a mesh configuration, and the matrix is inverted to yield the terminal impedance.

The terminal impedance is then expanded in partial fractions and partitioned into two matrices, one of which consists of elemental subnetworks that are readily realizable as "tuned" circuits in a star configuration. Necessary conditions for realizability of the remaining impedance are also derived in terms of properties of the numerator coefficient matrices and the residue matrices. After all elemental subnetworks are removed, the remaining impedance is irreducible in a star configuration, and the matrix is inverted again, thus returning to the terminal admittance.

The resulting network is realized canonically, as in Foster's procedure, with a "minimum" number of elements by avoiding partial pole removal and any partitioning that introduces excess nodes. The disadvantage of the star-mesh technique is that the realizable immittance matrix must have an equivalent network representation that has no "bridge" across internal nodes.

ON COUPLED OSCILLATORS*

J. K. AGGARWAL
C. G. RICHIE
Department of Electrical Engineering and Laboratories for Electronics and Related Science Research, The University of Texas
Austin, Texas

ABSTRACT

The behavior of coupled oscillators described by the differential equations

$$\ddot{x} + f(x,\dot{x}) + g(x,y) = 0$$

$$\ddot{y} + h(y,\dot{y}) + k(x,y) = 0$$

where $g(x,y) = U_x(x,y)$, $k(x,y) = U_y(x,y)$, and $U(x,y)$ is a positive definite potential function, has been investigated. Conditions on $f(x,\dot{x})$ and $h(y,\dot{y})$ have been derived which preclude the existence of sustained oscillations. For the case of coupled van der Pol oscillators, when $f(x,\dot{x}) = A_1\dot{x}^3 - B_1\dot{x}$ and $h(y,\dot{y}) = A_2\dot{y}^3 - B_2\dot{y}$, a Liapunov function of the form

$$V = \dot{x}^2/2 + \dot{y}^2/2 + U(x,y)$$

has been used to find a "cylinder" in the four-dimensional phase space such that no periodic solution may exist entirely inside or entirely outside the "cylinder." Also a subset of the space has been found from which all trajectories eventually leave and into which no trajectories may enter.

INTRODUCTION

Oscillations described by van der Pol's, Rayleigh's, and Duffing's equations have been extensively investigated. These are one-degree-of-freedom systems, and the reason behind intensive efforts on this class of systems is that analytical and topological methods may be applied to systems described by second order differential equations. However, for higher order equations, topological methods are not sufficiently developed, and consequently, results on systems with two or more degrees of freedom are wanting. The objective of this paper is to describe some properties of coupled oscillators, systems with several degrees of freedom.

Nonlinear two-degree-of-freedom conservative systems have been examined by Rosenberg and Atkinson [1] and Rosenberg [2,3]. Rosenberg has shown the existence of "normal modes" for conservative systems, and the knowledge of normal modes reduces the two-degree-of-freedom problem to two systems each with only a single degree-of-freedom. Also Rosenberg has considered a very special class of conservative n-degree-of-freedom systems in [4]. Bogoliubov and Mitropolsky [5] have considered monofrequency oscillations in weakly nonlinear systems with several degrees of freedom represented by a system of second order differential

---

*This work was supported jointly by the Joint Services Electronics Program under Grant AF - AFOSR 766 - 65, and by NSF Grant GK - 261.

equations. In general this technique is very laborious. Hale [6] has considered coupled van der Pol oscillations where the coupling is weak.

Ponzo [7] has used geometric methods in a study of oscillations described by the generalized van der Pol equation (a one-degree-of-freedom system). Ponzo has shown the existence of a region with the property that all trajectories enter the region, and no trajectories leave the region. Aggarwal [8,9] has investigated systems described by a pair of coupled first order differential equations using a similar technique. Aggarwal has found amplitude bounds, as well as other behavior of trajectories by considering regions in the phase space with confining properties. In this study the above mentioned one-degree-of-freedom method has been applied to systems with several degrees of freedom.

The equations for anode currents of two capacitively coupled anode oscillators are given by

$$\frac{d^2 i_1}{dt^2} - \left[a_1 - c_1\left(\frac{di_1}{dt}\right)^2\right]\frac{di_1}{dt} + w_1^2 i_1 + K_2 i_2 = 0;$$

$$\frac{d^2 i_2}{dt^2} - \left[a_2 - c_2\left(\frac{di_2}{dt}\right)^2\right]\frac{di_2}{dt} + w_2^2 i_2 + K_1 i_1 = 0$$

(see Nag [10]). Using suitable transformations the above equations may be written as

$$\ddot{x} + A(\dot{x}^3/3-\dot{x}) + Bx = -\alpha(x-y) \tag{1}$$

$$\ddot{y} + C(\dot{y}^3/3-\dot{y}) + Dy = \alpha(x-y). \tag{2}$$

In this paper, we shall consider the more general equations

$$\ddot{x} + f(x,\dot{x}) + g(x,y) = 0 \tag{3}$$

$$\ddot{y} + h(y,\dot{y}) + k(x,y) = 0 \tag{4}$$

where $g(x,y) = \frac{\partial U(x,y)}{\partial x}$, $k(x,y) = \frac{\partial U(x,y)}{\partial y}$, and $U(x,y)$ is a positive definite potential function. The condition that $g(x,y) = U_x(x,y)$ and $k(x,y) = U_y(x,y)$ where $U(x,y)$ is a potential function is not restrictive since this amounts to requiring the system

$$\ddot{x} + g(x,y) = 0$$

$$\ddot{y} + k(x,y) = 0$$

to be conservative. A typical example of $g(x,y)$ and $k(x,y)$ is

$$g(x,y) = \sum_{n}^{r} a_n x^n + \sum_{n}^{s} A_n (x-y)^n,$$

$$k(x,y) = \sum_{n}^{q} b_n y^n - \sum_{n}^{s} A_n (x-y)^n$$

where a_n, b_n, and A_n are positive for all n and n, r, s, q are odd integers. We shall assume the functions f, g, h and k to be well behaved (Lipshitz) to guarantee unique and continuous solutions to the differential equations.

ASYMPTOTIC BEHAVIOR

The system described by equations (3) and (4) has no sustained oscillatory solution and further, all trajectories tend to the origin provided that

$$\text{(i)} \quad \dot{x}f(x,\dot{x}) > 0 \text{ for } \dot{x} \neq 0, \text{ and } f(x,0) = 0$$

$$\text{(ii)} \quad \dot{y}h(y,\dot{y}) > 0 \text{ for } \dot{y} \neq 0, \text{ and } h(y,0) = 0.$$

The above result follows immediately from considering the Liapunov function $V = \dot{x}^2/2 + \dot{y}^2/2 + U(x,y)$ and using LaSalle's Theorem [11].

This result can be generalized to the case of systems described by n coupled second order differential equations of the form

$$\ddot{x}_i + f_i(x_i,\dot{x}_i) + g_i(x_1,x_2,\dots,x_n) = 0, \; i = 1,2,\dots,n.$$

Here $g_i(x_1,x_2,\dots,x_n) = U_{x_i}(x_1,x_2,\dots,x_n)$

with the restrictions that

$$\dot{x}_i f_i(x_i,\dot{x}_i) > 0 \text{ for } \dot{x}_i \neq 0, \; f_i(x_i,0) = 0,$$

and

$$U \to \infty \text{ as } \| x \| \to \infty,$$

where $\| x \| = (x_1^2 + x_2^2 + \dots + x_n^2)^{1/2}$. It may be observed that this result is not surprising since the system of equations describes a circuit with conservative capacitors and inductors and dissipative resistors. However, this result is a good introduction for the following section on coupled van der Pol oscillators.

COUPLED VAN DER POL OSCILLATORS

The equations for two coupled van der Pol oscillators

$$\ddot{x} + A(x^2-1)\dot{x} + Bx = -\alpha(x-y)$$

$$\ddot{y} + C(y^2-1)\dot{y} + Dy = +\alpha(x-y)$$

may be transformed to equations (1) and (2). It may be added that such a transformation is possible only if the coupling is linear. We shall consider the more general system

$$\ddot{x} + A_1\dot{x}^3 - B_1\dot{x} + U_x(x,y) = 0 \tag{5}$$

$$\ddot{y} + A_2\dot{y}^3 - B_2\dot{y} + U_y(x,y) = 0 \tag{6}$$

where $U_x(x,y)$ and $U_y(x,y)$ are as before and A_1, A_2, B_1 and B_2 are all positive.

For the system described by equations (5) and (6) no sustained oscillatory solution exists which is entirely confined in the "cylinder" defined by

$$A_1\dot{x}^4 + A_2\dot{y}^4 = B_1\dot{x}^2 + B_2\dot{y}^2.$$

Let $V = \dot{x}^2/2 + \dot{y}^2/2 + U(x,y)$, then it is readily seen that

$$\dot{V} = -\dot{x}(A_1\dot{x}^3 - B_1 x) - \dot{y}(A_2\dot{y}^3 - B_2\dot{y}).$$

Since V is a positive definite function and $\dot{V}$ is positive in the neighborhood of the origin, the origin is an unstable singular point. Consider the surface Γ where $\dot{V} = 0$, i.e.

$$A_1\dot{x}^4 + A_2\dot{y}^4 = B_1\dot{x}^2 + B_2\dot{y}^2.$$

Inside Γ, $\dot{V} > 0$; hence no periodic solution can exist entirely within Γ. Similarly, it may be shown that no periodic solution can be entirely confined outside the surface Γ.

The projection of this surface on the x - y plane is the closed curve γ in Figure 1 for $A_1 = 5/3$, $B_1 = 5$, $A_2 = 10/3$, $B_2 = 10$. Let r_0 be the radius of the circle of largest diameter σ which may be inscribed in the interior of the curve γ. Then

$$r_0^2 = \min.\left[\frac{B_2}{A_2},\frac{B_1}{A_1}\right].$$

It may be observed that trajectories starting inside the surface Σ given by

$$\frac{1}{2}r_0^2 = \frac{1}{2}\dot{x}^2 + \frac{1}{2}\dot{y}^2 + U(x,y)$$

leave the enclosed space and no trajectory can enter the enclosed space. This is true because for all values of x and y the circle

$$\dot{x}^2/2 + \dot{y}^2/2 = r_0^2/2 - U(x,y)$$

is inside the cylinder Γ.

NUMERICAL RESULTS

Figures 2, 3, and 4 show the solution to (1) and (2) for A = 5, B = 1, C = 10, D = 2, $\alpha = 0.5$ with initial conditions of x = 10, $\dot{x} = 2$, y = 10, and $\dot{y} = 2$. The solution was obtained by numerically integrating the equations on the CDC 1604 digital computer using the method of

Nordsieck [12]. The minimum value of V obtained for the above trajectory was 3.06, clearly outside the surface Σ given by $1.5 = \dot{x}^2/2 + \dot{y}^2/2 + x^2/2 + 0.25(y-x)^2$ in this instance. Note also that the solution reaches a periodic orbit; however, for small values the coupling coefficient $\alpha = 0.1$ the solutions are not periodic.

DISCUSSION

An inner bound for periodic and aperiodic oscillations has been derived, and currently work is being done to develop an outer bound. These bounds, as described above, are spherical in nature, a closed surface in the 2n dimensional phase space where n is the number of degrees of freedom. It may be observed that such a spherical type bound will not, in the limit, give the solution, since the solution is a curve in the 2n dimensional space and not a hypersurface. A limiting hypersurface (if one exists) is not a solution, since that would imply the intersection of solutions. This has already been excluded since f, g, h and k are Lipshitzian. This situation is most unlike the case in two dimensions where the limit of the bounds gives the solution. The corresponding type of surface in 3 or more dimensions would be a toroid since it tends to a curve in the limit as the cross-section decreases to zero. Spherical bounds do, however, give the subset of phase space where the toroid may be located.

The dependence of the existence of a periodic oscillation on coupling is interesting; for in the case of linear oscillations the existence of periodic oscillations depends on the periods of two oscillators being commensurate and is independent of the coupling.

BIBLIOGRAPHY

1. R. M. Rosenberg and C. P. Atkinson, "On Natural Modes and Their Stability in Nonlinear Two-Degree-of-Freedom Systems," J. Appl. Mech., Vol. 26, 1959, pp. 377-385.

2. R. M. Rosenberg, "Normal Modes of Nonlinear Dual-Mode Systems," J. Appl. Mech., Vol. 27, June, 1960, pp. 263-268.

3. R. M. Rosenberg, "On Normal Vibrations of a General Class of Nonlinear Dual-Mode Systems," J. Appl. Mech., June, 1961, pp. 275-283.

4. R. M. Rosenberg, "The Normal Modes of Nonlinear n-Degree of Freedom Systems," J. Appl. Mech., March, 1962, pp. 7-14.

5. N. N. Bogoliubov and Y. A. Mitropolsky, Asymptotic Methods in the Theory of Non-Linear Oscillations, Hindustan Publishing Corporation (India), 1961.

6. J. K. Hale, Oscillations in Nonlinear Systems, McGraw-Hill Book Company, Inc., pp. 165-169, 146-148.

7. P. J. Ponzo, "Nonlinear Oscillations in Second Order Systems," Coordinated Science Laboratory Report R-193, University of Illinois, Urbana, Illinois, June, 1964.

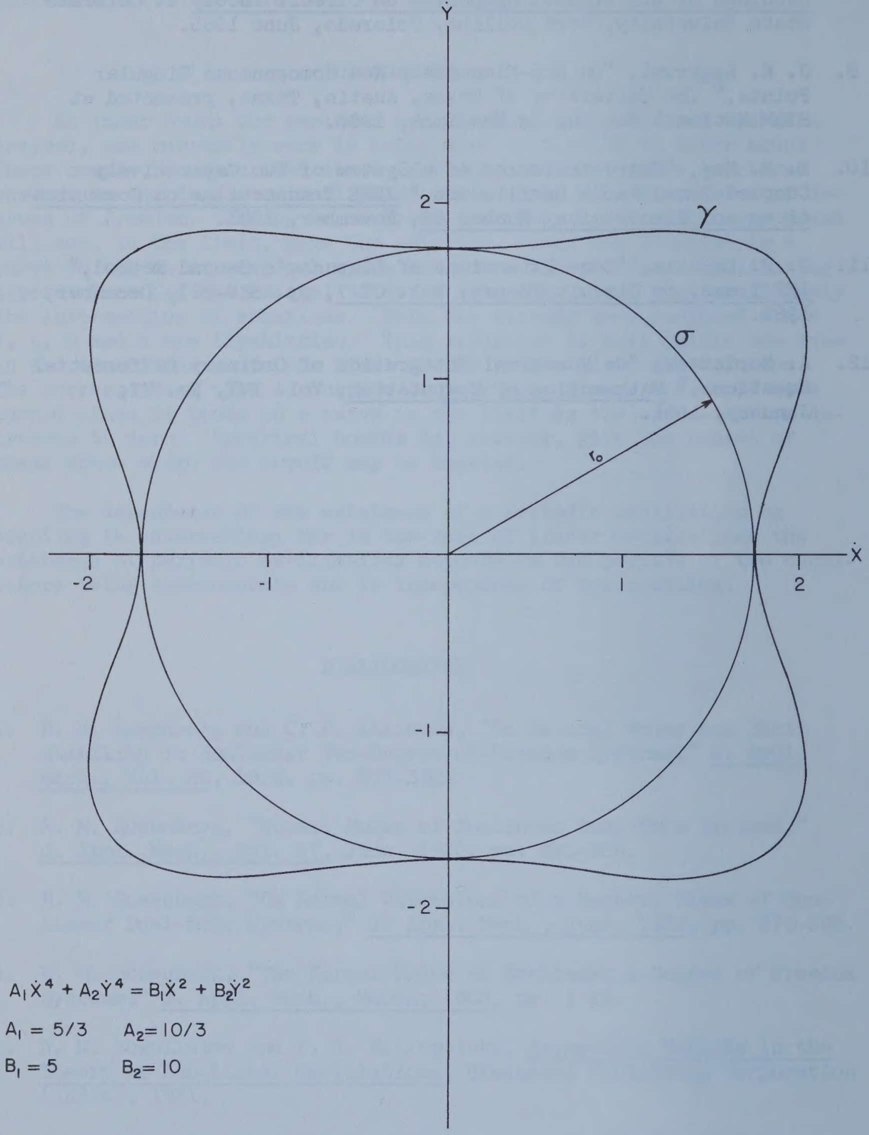
8. J. K. Aggarwal, "Nonlinear Oscillations: Amplitude Bounds," Proceedings of 8th Midwest Symposium on Circuit Theory at Colorado State University, Fort Collins, Colorado, June 1965.

9. J. K. Aggarwal, "On Non-Elementary Non Homogeneous Singular Points," The University of Texas, Austin, Texas, presented at SIAM National Meeting in New York, 1965.

10. B. R. Nag, "Characteristics of a System of Two Capacitively Coupled Tuned Anode Oscillators," IEEE Transactions on Communications and Electronics, Number 69, November, 1963.

11. J. P. LaSalle, "Some Extensions of Liapunov's Second Method," IRE Trans. on Circuit Theory, Vol. CT-7, pp. 520-527, December, 1960.

12. A. Nordsieck, "On Numerical Integration of Ordinary Differential Equations," Mathematics of Computation, Vol. XVI, No. 77, January, 1961.



$A_1\dot{X}^4 + A_2\dot{Y}^4 = B_1\dot{X}^2 + B_2\dot{Y}^2$

$A_1 = 5/3 \qquad A_2 = 10/3$

$B_1 = 5 \qquad B_2 = 10$

FIGURE I

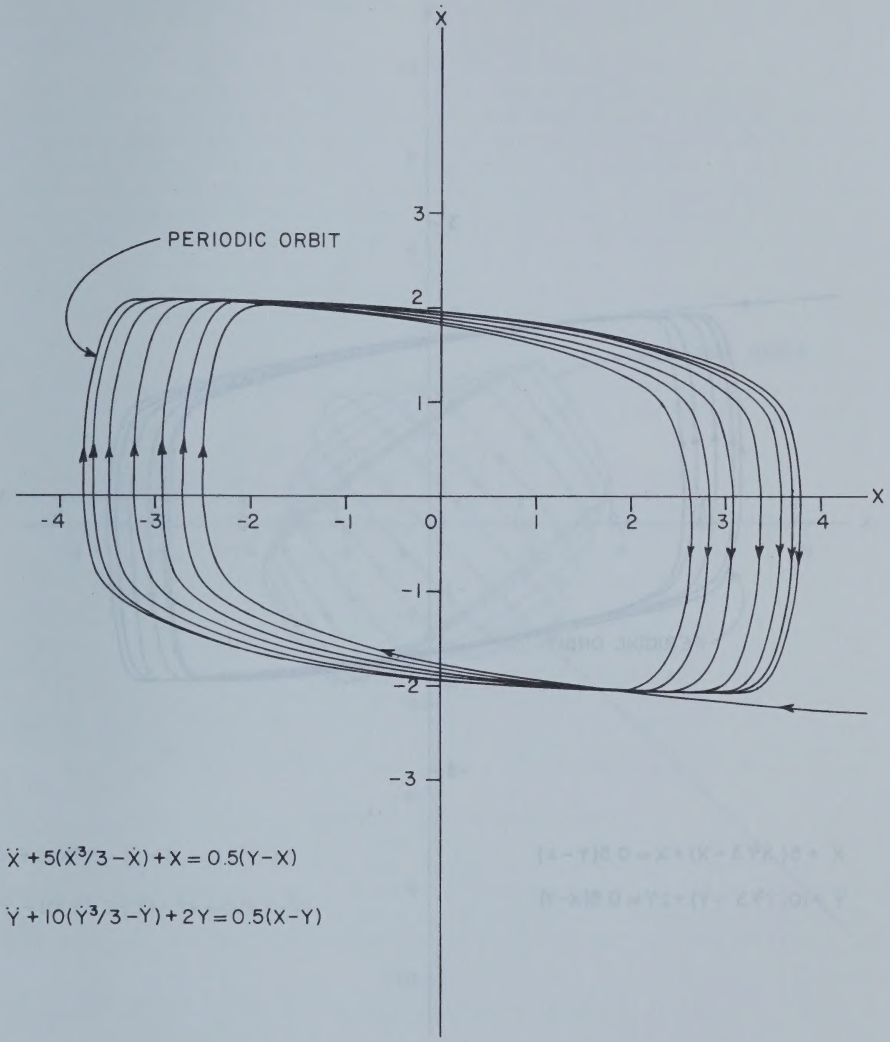
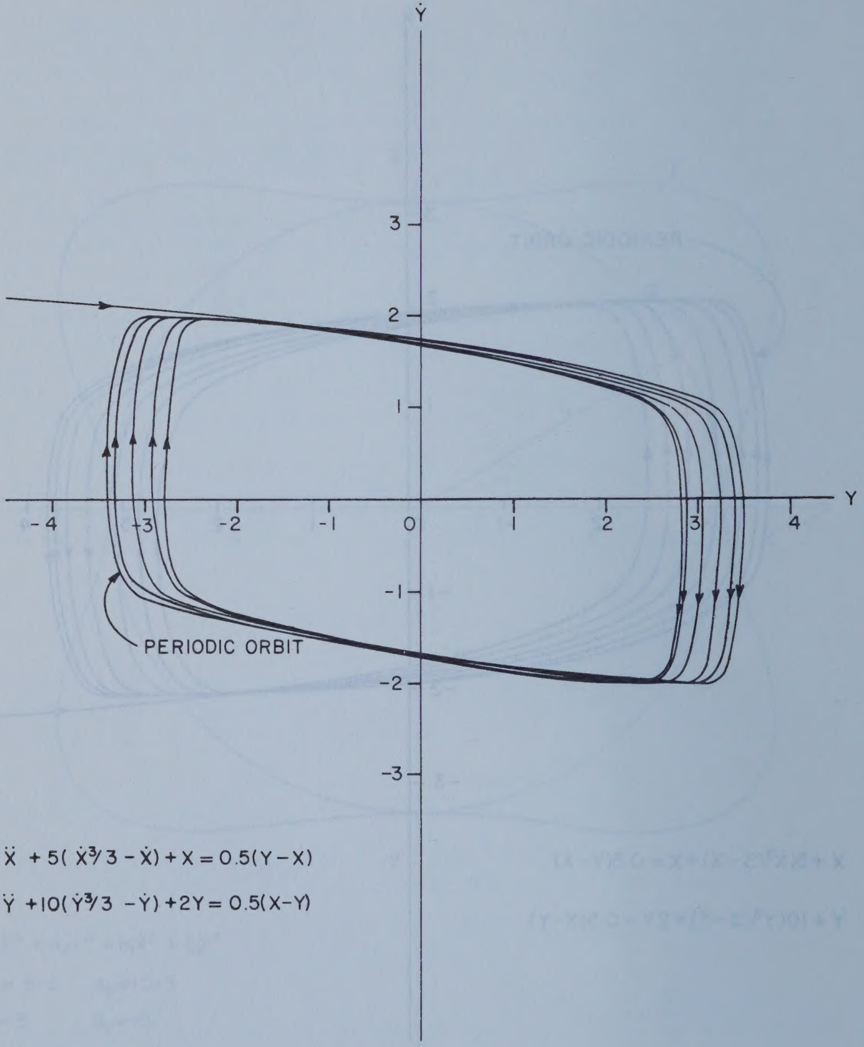

$\dot{X}$
3
2
1
-1
-2
-3
PERIODIC ORBIT
X
-4
-3
-2
-1
0
1
2
3
4
$\ddot{X} + 5(\dot{X}^3/3 - \dot{X}) + X = 0.5(Y - X)$
$\ddot{Y} + 10(\dot{Y}^3/3 - \dot{Y}) + 2Y = 0.5(X - Y)$

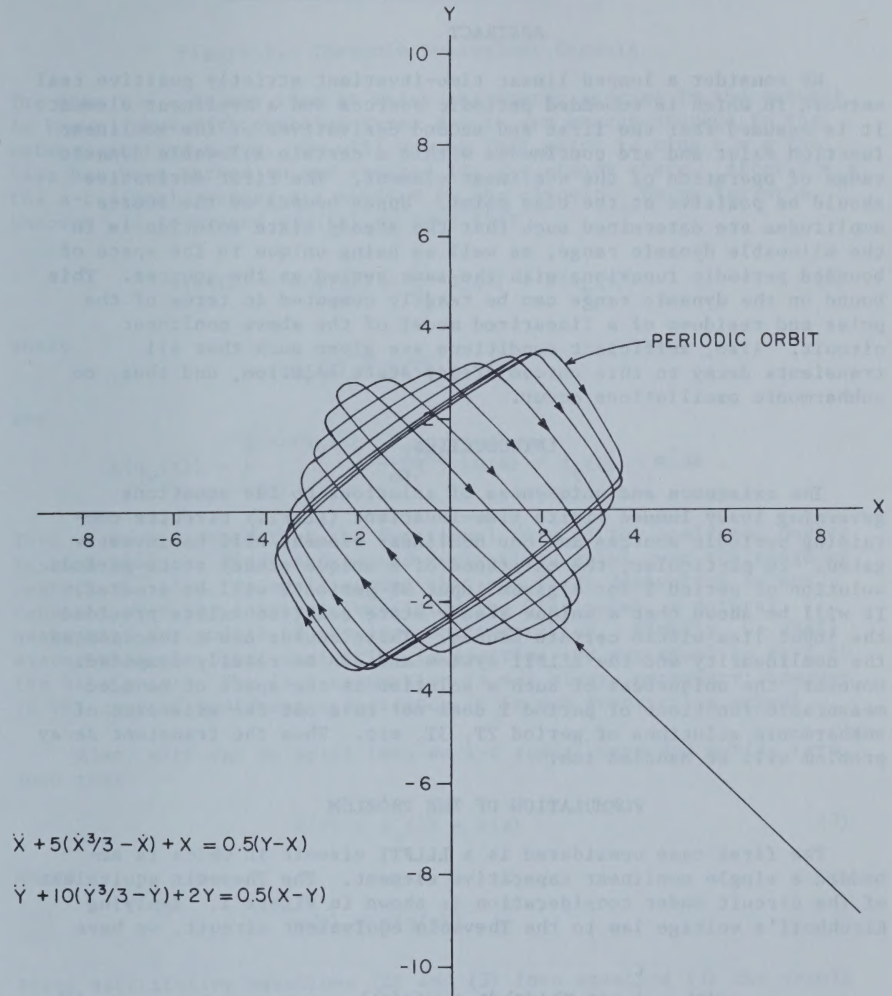
FIGURE 2



$$\ddot{X} + 5(\dot{X}^3/3 - \dot{X}) + X = 0.5(Y - X)$$

$$\ddot{Y} + 10(\dot{Y}^3/3 - \dot{Y}) + 2Y = 0.5(X - Y)$$

FIGURE 3



$$\ddot{X} + 5(\dot{X}^3/3 - \dot{X}) + X = 0.5(Y - X)$$

$$\ddot{Y} + 10(\dot{Y}^3/3 - \dot{Y}) + 2Y = 0.5(X - Y)$$

FIGURE 4

LOWER BOUNDS ON THE THRESHOLDS OF SUBHARMONIC OSCILLATIONS

T. N. TRICK
Department of Electrical Engineering
University of Illinois
Urbana, Illinois

D. R. ANDERSON
School of Electrical Engineering
Purdue University
Lafayette, Indiana

ABSTRACT

We consider a lumped linear time-invariant strictly positive real network in which is embedded periodic sources and a nonlinear element. It is assumed that the first and second derivatives of the nonlinear function exist and are continuous within a certain allowable dynamic range of operation of the nonlinear element. The first derivative should be positive at the bias point. Upper bounds on the source amplitudes are determined such that the steady state solution is in the allowable dynamic range, as well as being unique in the space of bounded periodic functions with the same period as the sources. This bound on the dynamic range can be readily computed in terms of the poles and residues of a linearized model of the above nonlinear circuit. Also, sufficient conditions are given such that all transients decay to this unique steady state solution, and thus, no subharmonic oscillations occur.

INTRODUCTION

The existence and uniqueness of solutions to the equations governing lossy lumped finite time-invariant (LLLFTI) circuits containing periodic sources and one nonlinear element will be investigated. In particular, the existence of a unique steady state periodic solution of period T for a given input of period T will be studied. It will be shown that a unique steady state solution exists provided the input lies within certain bounds. These bounds are a function of the nonlinearity and the LLLFTI system and can be readily computed. However, the uniqueness of such a solution in the space of bounded measurable functions of period T does not rule out the existence of subharmonic solutions of period 2T, 3T, etc. Thus the transient decay problem will be handled too.

FORMULATION OF THE PROBLEM

The first case considered is a LLLFTI circuit in which is embedded a single nonlinear capacitive element. The Thevenin equivalent of the circuit under consideration is shown in Figure 1. Applying Kirchhoff's voltage law to the Thevenin equivalent circuit, we have

$$e(t) = \int_a^t z(t-\tau)\, i(\tau)\, d\tau + v(q(t)) \quad . \tag{1}$$

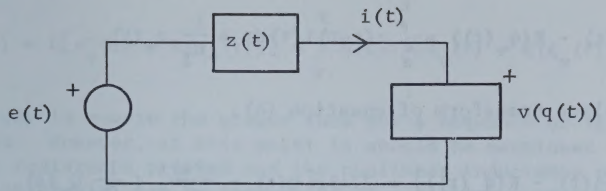


Figure 1. Thevenin Equivalent Circuit

The time t = a is any time at which the energy storage in the network is known, thus e(t) contains terms due to any energy storage in the network at time t = a. We will assume that prior to time t = a the bias has been turned on and reached its d-c steady state. After t = a the a-c signal sources are applied to the circuit. Using Taylor's theorem [1] to expand v(q(t)) we have that

$$v(q(t)) = v(q(a)) + \frac{1}{C_\ell} q_v(t) + R(q_v(t)) \tag{2}$$

where

$$q_v(t) = q(t) - q(a)$$

and

$$R(q_v(t)) = \int_{q(a)}^{q(a)+q_v(t)} \frac{d^2 v(\alpha)}{d\alpha^2} [q(a) + q_v(t) - \alpha] d\alpha .$$

Thus we have assumed that the functional form of the nonlinear device is such that the first and second derivatives exist over the range of operation. Also, we shall assume that the first derivative is not constant over any interval containing the bias point, a natural assumption for models of physical nonlinearities. Finally, we shall assume that the first derivative is positive and not equal to zero at the bias point. The last assumption is not always necessary; however, in the case of nonlinear capacitance it is not restrictive anyway.

Also, e(t) can be split into an a-c signal term and a bias term such that

$$e(t) = e_v(t) + e(a) \tag{3}$$

where

$$e(a) = v(q(a)) .$$

After substituting equations (2) and (3) into equation (1) the result is that

$$e_v(t) = \int_a^t z(t-\tau) i(\tau) d\tau + \frac{1}{C_\ell} q_v(t) + R(q_v(t)) ,$$

or

$$e_v(t) - R(q_v(t)) = \int_a^t z(t-\tau)\, i(\tau)\, d\tau + \frac{1}{C_\ell} q_v(t) \; . \tag{4}$$

Taking the Laplace transform of equation (4),

$$\mathcal{L}\{e_v(t) - R(q_v(t))\} = Z(s)\, s\left[Q(s) - \frac{q(a)}{s}\right] + \frac{1}{C_\ell} Q_v(s) ,$$

but

$$Q(s) - \frac{q(a)}{s} = Q_v(s)$$

so that

$$\mathcal{L}\{e_v(t) - R(q_v(t)) = \left\{sZ(s) + \frac{1}{C_\ell}\right\} Q_v(s) \; . \tag{5}$$

Solving for $Q_v(s)$,

$$Q_v(s) = \left\{\frac{Y(s)}{s}\right\} \mathcal{L}\,\{e_v(t) - R(q_v(t))\} \tag{6}$$

where

$$Y(s) = \frac{1}{Z(s) + \dfrac{1}{C_\ell s}} \; .$$

Note that $Y(s)$ has a zero at $s = 0$ and cannot have a pole at $s = 0$ due to the series capacitance. However, $Y(s)$ can have a zero at $s = \infty$ (series inductance which is always present in the form of lead inductance), or, if series inductance is neglected, $Y(s)$ can have a pole at $s = \infty$ which means a capacitance, C_s, shunts the nonlinearity. In any case, the numerator $Y(s)/s$ cannot have a degree greater than the denominator, and $Y(s)/s$ can have only zeros at $s = \infty$. If it is assumed that $Z(s)$ is a strictly positive real function*, then it is obvious that the poles of $Y(s)/s$ must be strictly in the left-half plane. The function $Y(s)/s$ can be expanded into

$$\frac{Y(s)}{s} = C + H(s) \tag{7}$$

where $C = C_\ell C_s / C_\ell + C_s$, and $H(s)$ is the Laplace transform of an absolutely integrable function since the poles of $Y(s)/s$ are strictly in the left-half space. Hence, taking the inverse Laplace transform of equation (6),

---

*A function $Z(s)$ is said to be strictly P.R. if $\mathrm{Re}\, Z(s) > 0$ for all $\mathrm{Re}\, s \geq 0$ (excluding poles or zeros at the origin and at infinity), and if $Z(s)$ is real for s real.

$$q_v(t) = C[e_v(t) - R(q_v(t))] + \int_a^t h(t-\tau)[e_v(\tau) - R(q_v(\tau))]d\tau. \quad (8)$$

Equation (8) is now in the proper form for a sequence of iterates to be defined. However, at this point it should be mentioned that the nonlinear resistance problem and the nonlinear inductance problem, without hysteresis, can be formulated in a similar manner [2].

AN ESTIMATE OF THE BOUND ON THE DYNAMIC RANGE

Once the problem has been formulated it is a simple matter to define a sequence of iterates. The question is whether or not these iterates converge to a unique solution. Since we are looking for the steady state solution we have that

$$q_v(t) = C[e_v(t) - R(q_v(t))] + \int_{-\infty}^t h(t-\tau)[e_v(\tau) - R(q_v(\tau))]d\tau \quad (9)$$

or

$$q_v(t) = C[e_v(t) - R(q_v(t))] + \sum_{n=0}^{\infty} \int_{t-(n+1)T}^{t-nT} h(t-\tau)[e_v(\tau)-R(q_v(\tau))]d\tau. \quad (10)$$

Let $\lambda = t-nT-\tau$, then

$$q_v(t) = C[e_v(t)-R(q_v(t))] + \sum_{n=0}^{\infty} \int_0^T h(\lambda+nT)[e_v(t-\lambda-nT) -R(q_v(t-\lambda-nT))]d\lambda. \quad (11)$$

From the previous formulation of the problem it is clear that summation and integration may be interchanged in eq. (11), and making use of the periodicity of $e_v(t)$ and $q_v(t)$ results in the following equation.

$$q_v(t) = C[e_v(t)-R(q_v(t))] - \int_0^T \sum_{n=0}^{\infty} h(\lambda+nT)[e_v(t-\lambda) -R(q_v(t-\lambda))]d\lambda. \quad (12)$$

The sequence of iterates is defined as

$$q_v^1(t) = C\, e_v(t) - \int_0^T h_T(\lambda) e_v(t-\lambda)\, d\lambda \quad (13)$$

and

$$q_v^n(t) = C[e_v(t)-R(q_v^{n-1}(t))] - \int_0^T h_T(\lambda)[e_v(t-\lambda) -R(q_v^{n-1}(t-\lambda))]d\lambda \quad (14)$$

where

$$\sum_{n=0}^{\infty} h(\lambda+nT) = h_T(\lambda).$$

The justification for assuming the above form for the iterates is that for $e_V(t)$ small enough we know that $q_V(t)$ is periodic with period T [3]. We would like to know how large the amplitude of $e_V(t)$ can be such that the solution $q_V(t)$ remains periodic of period T.

If there exists a positive number $d < 1$ such that

$$\|q_V^{n+1}(t) - q_V^{n}(t)\| \leq d\, \|q_V^{n}(t) - q_V^{n-1}(t)\|, \text{ for all } n,$$

and where

$$\|q_V(t)\| = \sup_{\text{all } t} |q_V(t)|,$$

then the above sequence of iterates converges, and

$$\|q_V(t)\| \leq \frac{\|q_V^{1}(t)\|}{1-d} = Q \tag{15}$$

One can readily show that

$$d = \bar{S}(Q)\ \{C + \int_0^T |h_T(\lambda)|\, d\lambda\}, \tag{16}$$

where

$$v'(q(t)) - v'(q(a)) = R'(q_V(t)),$$

and thus,

$$\bar{S}(Q) = \max_{\|q_V(t)\| \leq Q} |v'(q(t)) - v'(q(a))|.$$

We must find a solution to equation (16) such that $d < 1$. If such a d can be found, then the steady state solution is unique in the space of bounded measurable periodic functions with period T.

It is clear that such a solution for d exists since $\bar{S}(0) = 0$, and $\bar{S}(Q)$ is a continuous function. Hence, remembering that

$$Q = \frac{\|q_V^{1}(t)\|}{1-d},$$

we see that by making $\|q_V^{1}(t)\|$ small enough, we can find a $d < 1$. However, this is not enough. One would like to find the largest possible value of $\|q_V^{1}(t)\|$, and thus $\|e_V(t)\|$, such that $d < 1$. This can be done by graphical or analytical techniques, or by trial and error. In the abrupt junction varactor diode the value of d which yields the maximum $\|e_V(t)\|$ is d = 1/2.

It should be noted that

$$\int_0^T |h_T(\lambda)|\,d\lambda = \int_0^T \Big|\sum_{n=0}^{\infty} h(\lambda+nT)\Big|\,d\lambda$$

$$\leq \sum_{n=0}^{\infty} \int_0^T |h(\lambda+nT)|\,d\lambda = \sum_{n=0}^{\infty} \int_{nT}^{(n+1)T} |h(\tau)|\,d\tau$$

$$= \int_0^{\infty} |h(\tau)|\,d\tau . \tag{17}$$

Thus our dynamic range estimate is better than a previous estimate [3] in the sense that a larger Lipschitz constant, $\bar{S}(Q)$, is tolerated, and so, a bigger range of operation of the nonlinear device is allowed.

A MORE READILY COMPUTABLE DYNAMIC RANGE ESTIMATE

The computation of the dynamic range estimate is straightforward. The only difficulty lies in the evaluation of the quantity

$$\int_0^T |h_T(\lambda)|\,d\lambda \ .$$

We will now show that the above integral can be easily estimated in terms of the poles and residues of $|H(s)|^2$ and the period T. A well known result from functional analysis is that

$$\frac{1}{T}\int_0^T |h_T(\lambda)|\,d\lambda \leq \Big(\frac{1}{T}\int_0^T |h_T(\lambda)|^2 d\lambda\Big)^{1/2} . \tag{18}$$

where $\int_0^T |h_T(\lambda)|^2 d\lambda$ obviously exists from the way the problem was formulated. In fact $h_T(\lambda)$ is continuous over the interval [0,T]. One can show that the right side of inequality (18) asymptoticly approaches the left side as T → 0. Rewriting equation (18) we have that

$$\int_0^T |h_T(\lambda)|\,d\lambda \leq \Big(T\int_0^T |h_T(\lambda)|^2 d\lambda\Big)^{1/2}.$$

Let us expand $h_T(\lambda)$ in a Fourier series over the interval [0,T]. Thus, we can write

$$h_T(\lambda) = \frac{1}{T}\sum_{m=-\infty}^{\infty} C_m e^{jm\frac{2\pi}{T}\lambda}$$

where the C_m's are simply the Fourier coefficients of $h_T(\lambda)$. Therefore,

$$T\int_0^T |h_T(\lambda)|^2 d\lambda = \frac{T}{T^2}\int_0^T \left| \sum_{m=-\infty}^{\infty} C_m e^{jm\frac{2\pi}{T}\lambda} \right|^2 d\lambda$$

$$= \sum_{m=-\infty}^{\infty} |C_m|^2$$

and so

$$\int_0^T |h_T(\lambda)| d\lambda \leq (T\int_0^T |h_T(\lambda)|^2 d\lambda)^{1/2} = \left(\sum_{m=-\infty}^{\infty} |C_m|^2 \right)^{1/2} .$$

But the coefficient

$$C_m = \int_0^T h_T(\lambda) e^{-jm\frac{2\pi}{T}\lambda} d\lambda = H(jm\omega_0) \qquad (19)$$

where $\omega_0 = 2\pi/T$, and $H(jm\omega_0)$ is the Fourier transform of $h(t)$ evaluated at the discrete frequencies $jm\omega_0$. Thus

$$\int_0^T |h_T(\lambda)| d\lambda \leq \left(\sum_{m=-\infty}^{\infty} |H(jm\omega_0)|^2 \right)^{1/2} . \qquad (20)$$

The problem of evaluating $\int_0^T |h_T(\lambda)| d\lambda$ has been reduced to evaluating the square root of an infinite sum. This may not seem like a simplification, however, this infinite sum can be put into closed form in terms of the poles and residues of $|H(s)|^2$ resulting in a quick and easy dynamic range estimate.

First let us consider the case where

$$H(s) = \frac{a_\ell s^\ell + a_{\ell-1} s^{\ell-1} + \ldots + a_1 s + a_0}{s^k + b_{k-1} s^{k-1} + \ldots + b_1 s + b_0} \qquad (21)$$

and all the poles of $H(s)$ are simple and strictly in the left-half plane, and k is always greater than ℓ. Then

$$H(s)H(-s) = |H(s)|^2$$

$$= \frac{a_\ell s^\ell + \ldots a_1 s + a_0}{(s+s_1)(s+s_2)\ldots(s+sk)} \times \frac{a_\ell (-s)^\ell + \ldots - a_1 s + a_0}{(-s+s_1)(-s+s_2)\ldots(-s+s_k)}$$

$$= \frac{K_1}{s+s_1} + \frac{K_2}{s+s_2} + \ldots + \frac{K_k}{s+s_k} + \frac{K_1'}{-s+s_1} + \frac{K_2'}{-s+s_2} + \ldots + \frac{K_\ell'}{-s+s_k} \qquad (22)$$

where $s_1, \ldots s_k$ are the poles of H(s) and $K_1, K_1', \ldots, K_k, K_k'$ are the residues. Note that

$$K_1 = K_1', \; K_2 = K_2', \; \ldots, \; K_k = K_k'$$

so that

$$|H(s)|^2 = \frac{2s_1K_1}{-s^2+s_1^2} + \frac{2s_2K_2}{-s^2+s_2^2} + \ldots + \frac{2s_kK_k}{-s^2+s_k^2} \; . \tag{23}$$

If we let $s = jm\omega_0$, we get

$$\sum_{m=-\infty}^{\infty} |H(jm\omega_0)|^2 = \sum_{m=-\infty}^{\infty} \left\{ \frac{2s_1K_1/\omega_0^2}{m^2+s_1^2/\omega_0^2} + \ldots + \frac{2s_kK_k/\omega_0^2}{m^2+s_k^2/\omega_0^2} \right\}$$

$$= \frac{2K_1}{s_1} + \ldots + \frac{2K_k}{s_k} + 2\sum_{m=1}^{\infty} \left\{ \frac{2s_1K_1/\omega_0^2}{m^2+s_1^2/\omega_0^2} + \ldots + \frac{2s_kK_k/\omega_0^2}{m^2+s_k^2/\omega_0^2} \right\} . \tag{24}$$

However, it is known [4] that

$$\sum_{n=1}^{\infty} \frac{1}{-n^2+z^2} = \frac{\pi}{2z} \cot \pi z - \frac{1}{2z^2} \; . \tag{25}$$

Thus equation (24) can now be put into closed form and simplified so that

$$\sum_{m=-\infty}^{\infty} |H(jm\omega_0)|^2 = \frac{j2\pi K_1}{\omega_0} \cot \pi j \frac{s_1}{\omega_0} + \ldots + \frac{j2\pi K_k}{\omega_0} \cot \pi j \frac{s_k}{\omega_0}$$

or

$$\left(\sum_{m=-\infty}^{\infty} |H(jm\omega_0|^2 \right)^{1/2} = \left(\frac{j2\pi K_1}{\omega_0} \cot \pi j \frac{s_1}{\omega_0} + \ldots + \frac{j2\pi K_k}{\omega_0} \cot \pi j \frac{s_k}{\omega_0} \right)^{1/2} \tag{26}$$

It is important to note in formula (26) that the residues K_i, for $i = 1,2,\ldots,k$, are not the residues of H(s), but $K_i = \bar{K}_i H(-s_i)$ where $\bar{K}_i$ is the residue of H(s) at the pole s_i.

Multiple order poles of H(s) can be handled also. Suppose that

$$H(s) = \frac{P(s)}{(s+b)^r} \tag{27}$$

where r is an integer, b is a complex number, and the degree of the denominator is greater than the degree of the numerator. The quantity

$$|H(s)|^2 = \frac{P(s)}{(s+b)^r} \times \frac{P(-s)}{(-s+b)^r}$$

$$= \frac{K_r}{(s+b)^r} + \frac{K_{r-1}}{(s+b)^{r-1}} + \dots + \frac{K_1}{s+b} + \frac{K'_r}{(-s+b)^r}$$

$$+ \frac{K'_{r-1}}{(-s+b)^{r-1}} + \dots + \frac{K'_1}{(-s+b)} \ . \tag{28}$$

Again $K_r = K'_r$, $K_{r-1} = K'_{r-1}$, ..., $K_1 = K'_1$, and

$$\sum_{m=-\infty}^{\infty} |H(jm\omega_0)|^2 = \sum_{m=-\infty}^{\infty} \left\{ \frac{K_r/(j\omega_0)^r}{(m-jb/\omega_o)^r} + \dots + \frac{K_2/(j\omega_0)^2}{(m-jb/\omega_0)^2} \right.$$

$$\left. + \frac{K_r/(-j\omega_0)^r}{(m-b/j\omega_0)^r} + \dots + \frac{K_2/(-j\omega_0)^2}{(m-b/j\omega_0)^2} + \frac{2bK_1/(j\omega_0)^2}{-m^2+b^2/(j\omega_0)^2} \right\} . \tag{29}$$

The above summation for the multiple root case can also be put in closed form [4] so that

$$\left(\sum_{m=-\infty}^{\infty} |H(jm\omega_0)|^2 \right)^{1/2} = \left\{ \frac{K_r}{(j\omega_0)^r} \frac{\pi^2}{(r-1)!} \frac{\partial^{(r-2)}}{\partial\theta_1^{(r-2)}} \operatorname{cosec}^2 \pi\theta_1 \right.$$

$$+ \dots + \frac{K_2}{(j\omega_0)^2} \pi^2 \operatorname{cosec}^2 \pi\theta_1 + \frac{K_r}{(-j\omega_0)^r} \frac{\pi^2}{(r-1)!} \frac{\partial^{(r-2)}}{\partial\theta_2^{(r-2)}} \operatorname{cosec}^2 \pi\theta_2$$

$$\left. + \dots + \frac{K_2}{(-j\omega_0)^2} \pi^2 \operatorname{cosec}^2 \pi\theta_2 + \frac{2K_1\pi}{j\omega_0} \cot \frac{\pi b}{j\omega_0} \right\}^{1/2} , \tag{30}$$

where $\theta_1 = jb/\omega_0$ and $\theta_2 = b/j\omega_0$. Obviously, if the system has a mixture of simple poles and multiple poles, then the above results for both the simple pole case and the multiple pole case can be applied to this situation.

Now the dynamic range estimate is

$$d = \bar{S}(Q)\{C + (\sum_{m=-\infty}^{\infty} |H(jm\omega_0|^2)^{1/2}\} . \tag{31}$$

For a lumped system the function $|H(s)|^2$ can be expanded by partial fractions, and thus the quantity

$$\left(\sum_{m=-\infty}^{\infty} |H(jm\omega_0|^2 \right)^{1/2}$$

can be readily computed by means of equations (26) and (30).

EXPERIMENTAL RESULTS

In order to determine just how good these estimates are, a circuit which is known to have non-unique solutions, i.e., subharmonic solutions, was investigated [2]. The circuit is illustrated in Figure 2. It is well known that this simple circuit can support subharmonic oscillations of order one-half the frequency of the pump, if a linear approximation to the above circuit is resonant in the neighborhood of one-half the frequency of the pump [5,6].

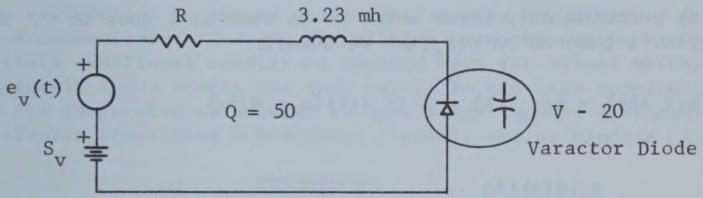


Figure 2: Subharmonic Oscillator

Applying the dynamic range estimate to the above circuit is was found that the bound on the pump voltage below which a unique steady state solution with period T existed was 5.85 volts rms. However, in the experiment it was found that the threshold at which one-half harmonics commenced was 4.8 volts rms. Thus, unfortunately, the existence of a unique steady state solution of period T, the period of the sources, in the space of bounded measurable functions of period T does not rule out the possibility of solutions in the space of bounded functions with period 2T, 3T, etc. In other words, if the steady state solution is unstable then, indeed, non-unique subharmonic oscillations could exist within the above bounds on the input amplitudes.

A solution to this stability problem can be obtained by finding bounds on the input within which all transients decay to the unique steady state solution of interest. This can be done as follows. The solution to the steady state problem is

$$q_p(t) = \int_{-\infty}^{t} h(t-\tau)[e_v(\tau)-R(q_p(\tau))\,d\tau + C[e_v(t)-R(q_p(t))] , \qquad (32)$$

and the solution including transients is

$$q_t(t) = \int_a^t h(t-\tau)[e_v(\tau)-R(q_t(\tau))]d\tau$$

$$+ C[e_v(t)-R(q_t(t))] . \qquad (33)$$

Taking the difference between eqs. (33) and (32) we have

$$q_t(t)-q_p(t) = \delta q = M(t) + \int_a^t h(t-\tau)[R(q_t(\tau))-R(q_p(\tau))]d\tau$$

$$+ C[R(q_t(t))-R(q_p(\tau))] , \qquad (34)$$

where M(t) contains only terms which go to zero as t goes to ∞. Again using Taylor's theorem on $R(q_t(t))$ we obtain

$$R(q_t(t)) = R(q_p(t)) + R'(q_p(t))\delta q + N(\delta q) \qquad (35)$$

where

$$N(\delta q) = \int_{q_p+q(a)}^{q_p+q(a)+\delta q} R''(\alpha)[q_p+q(a)+\delta q-\alpha]d\alpha.$$

Thus,

$$\delta q - C R'(q_p(t))\delta q - \int_a^t h(t-\tau)R'(q_p(\tau))\delta q \, d\tau$$

$$= M(t) + C N(\delta q) + \int_a^t h(t-\tau)N(\delta q)d\tau . \qquad (36)$$

We are assuming that $q_p(t)$ is known from the previous results, and we wish to show that δq goes to zero as t goes to infinity.

Going through the usual iteration procedure one obtains three sufficient conditions on the circuit in order that $\delta q \to 0$ as $t \to \infty$. The quantity δq is uniformly bounded by M, where

$$\|\delta q\| \leq \frac{\|M(t)\|}{1-d_3} = \bar{M} , \qquad (37)$$

provided that

(1) $d_3 = \dfrac{d_2}{1-d_1} < 1$,

(2) $d_2 = \bar{N}[C + \int_0^\infty |h(t)|dt] < 1$,

$$\text{where } \bar{N} = \max_{|\delta q| \leq \bar{M}} |R'(q(t)) - R'(q_p(t)+q(a))| ,$$

and

$$(3)\ d_1 = \bar{S}(\|q_p\|)[C + \int_0^\infty |h(t)|dt] < 1 .$$

The above three conditions can be satisfied by making $\|e_V(t)\|$ "small enough", and so, from eq. (37) we see that $\delta q \to 0$ as $t \to \infty$. No calculations have been made yet, however, it is believed that the above estimate will be better than a previous transient decay estimate [3], since the rate of decay of the transients above is not determined. Only the fact that the transients do decay to the unique steady state solution is determined.

CLOSING REMARKS

The restriction to lumped parameter circuits is primarily a matter of convenience. Indeed, distributed networks can be handled with certain additional conditions imposed upon the linear networks [3]. Actually these conditions turn out to be not very restrictive. Usually the computational problem is just slightly more difficult. Also, circuits containing N nonlinear elements can be handled [2].

REFERENCES

1. Boas, R. P., Jr., A Primer of Real Functions, John Wiley and Sons, Inc., New York, 1960.

2. Trick, T. N., "Lower Bounds on the Thresholds of Subharmonic Oscillations", Ph.D. Thesis, School of Electrical Engineering, Purdue University, 1966.

3. Leon, B. J. and Anderson, D. R., "The Stability of Pumped Nonlinear Reactance Circuits", IEEE Trans. on Circuit Theory, Vol. CT-10, No. 4, pp. 468-476, Dec. 1963. (A correction in Vol. CT-12, p. 143, March 1965.)

4. Jolley, L.B.W., Summation of Series, Dover Publications, Inc., New York, 1961.

5. Aiken, R. T. and Finzi, L. A., "Large Signal Analysis of a Parametric Second-Subharmonic Oscillator", IEEE Trans. on Circuit Theory, Vol. CT-10, No. 4, pp. 476-486, Dec. 1963.

6. Lytal, K. E., "Steady State Responses of Driven Networks Containing a Single Type Lossless Nonlinearity", Ph.D. Thesis, Dept. of Electrical Engineering, Purdue University, 1962.

ON THE DYNAMICAL EQUATIONS OF NONLINEAR NETWORKS WITH n-COUPLED ELEMENTS

CARLOS A. HOLZMANN
University of Notre Dame, Notre Dame, Indiana.
RUEY-WEN LIU*
University of California, Berkeley, California.

ABSTRACT

In this paper the existence of the normal form, for nonlinear networks with n-coupled elements described by characteristic parameters, is considered. Two choices of the dependent variables of the normal form are made: (i) the cut-set of charges, q, and the loop fluxes, Ψ; (ii) a subset of the characteristic parameters in the inductive and capacitive elements. In each case the sufficient conditions for the existence of the normal form are given, with respect to any normal distribution of independent sources, such that: (i) the normal form is defined in the entire space of the dependent variables, (ii) the set of dependent variables form a complete set which is also a state of the network.

I. INTRODUCTION

It is well known that not all nonlinear networks can be described by the normal form of the differential equations [7,9]. Efforts have been made to find the conditions on the network such that its dynamical behaviour can be described by the normal form [1-10]. The importance of this description is based on the following facts: (i) it is adaptable to almost all existing mathematical theories about nonlinear ordinary differential equations, (ii) it is suitable for computer programming, and (iii) a state of the network can be easily found.

Networks with n-coupled elements have been considered in [7,10]** In this paper we will approach the problem by combining the techniques suggested in [2,6,7,8].

II. PARAMETRIC REPRESENTATION OF n-COUPLED ELEMENTS

The parametric representation of a non-coupled (time-varying) two terminal resistor is given as $v(t)=V(z(t),t)$ and $i(t)=I(z(t),t)$ [8]. If the resistor is current-controlled (i-controlled) [6] then we have $v(t)=V(i(t),t)$ and $i(t)=(t)$, which can be obtained by letting $z=i$ and $I(z,t)=z$. Similarly, the characterization of a voltage-controlled (v-controlled) resistor, is obtained as $i(t)=I(v(t),t)$ and $v(t)=v(t)$, by letting $z=v$ and $V(z,t)=z$.

As an extension, we shall consider the type of n-coupled two-terminal elements which can be represented in the parametric form

$$\begin{aligned} v_k(t) &= V_k(z_1(t),\dots z_n(t);t) \\ i_k(t) &= I_k(z_1(t),\dots z_n(t);t) \quad , \quad k=1,\dots n \end{aligned} \tag{1'}$$

---

*Leave of absence from the University of Notre Dame, 1965-1966.

**It has come to our attention that E. S. Kuh will present a paper at the N.E. Conference, Chicago, October, 1965, covering related topics.

We shall say that the k-th element is voltage-controlled if it can be represented by $V_k(z_1 \ldots z_n, t) = z_k$, and current-controlled if by $I_k(z_1 \ldots z_n; t) = z_k$.

This representation includes many physical elements. For example, a vacuum triode in the grounded cathode configuration can be described by two mutually coupled resistors whose parametric representation can be obtained as $v_{gk} = z_1$, $i_g = I_g(z_2)$, $v_{pk} = z_2$, and $i_p = I_p(z_1, z_2)$, where $I_g(\cdot,\cdot)$ and $I_p(\cdot,\cdot)$ represent the grid and plate characteristics of the tube. We observe that one element is v_{gk}-controlled and the other is v_{pk}-controlled. A transistor triode in the common emitter configuration can be represented by $v_{be} = V_{be}(z_1, z_2)$, $i_b = z_1$, $v_{ce} = z_2$, and $i_c = I_c(z_1, z_2)$, where $V_{be}(\cdot,\cdot)$ and $I_c(\cdot,\cdot)$ are the base and collector characteristics of the transistor, respectively. One element is i_b-controlled and the other is v_{ce}-controlled.

A voltage amplifier can be modeled as $v_1 = z_1$, $i_1 = I_1(z_1)$, which is a v-controlled input, and $v_2 = V_2(z_1, z_2)$, $i_2 = z_2$, which is an i-controlled output.

Similarly, a dependent voltage source can be modeled as $v_1 = z_1$, $i_1 = 0$, which is a v-controlled input, and $v_2 = V_2(z_1, z_2)$, $i_2 = z_2$, which is i-controlled output. If it is "ideal" then $v_2 = V_2(z_1)$.

Equation (1') can be put into vector form as

$$\underline{v}_R(t) = \underline{V}_R(\underline{z}(t), t) \quad , \quad \underline{i}_R(t) = \underline{I}_R(\underline{z}(t), t) \tag{1}$$

Similarly, the parametric representation of n-coupled two-terminal capacitors and inductors is given by

$$\underline{v}_C(t) = \underline{V}_C(\underline{x}(t), t) \quad , \quad \underline{q}_C(t) = \underline{Q}_C(\underline{x}(t), t) \tag{2}$$

$$\underline{\psi}_L(t) = \underline{\phi}_L(\underline{y}(t), t) \quad , \quad \underline{i}_L(t) = \underline{I}_L(\underline{y}(t), t) \tag{3}$$

in which the voltage- and charge-controlled capacitor (v- and q-controlled), and flux- and current-controlled inductor (ψ- and i-controlled), are defined as in the case of a resistor.

Equations (1), (2) and (3) are called the characteristic equations of n-coupled resistors, capacitors, and inductors respectively. They give the $(\underline{v}_R, \underline{i}_R)$, $(\underline{v}_C, \underline{q}_C)$, and $(\underline{\psi}_L, \underline{i}_L)$ relationships of the respective elements. The variables $\underline{v}_R, \underline{i}_R, \underline{v}_C, \underline{q}_C, \underline{\psi}_L$, and $\underline{i}_L$ are called the branch characteristic variables, and $\underline{x}, \underline{y}, \underline{z}$ are called the characteristic parameters.

Note that we may include all of those n-port elements which can be replaced by n-coupled two-terminal elements. Obviously, the case n=1 includes the conventional C,R,L network elements. The case n=2 includes transistors, vacuum tubes, dependent sources, transformers, etc.

III. DYNAMICAL EQUATIONS

a. Characterization of the Network

We shall consider networks N consisting of n-coupled time-varying elements of the capacitive type C, of the resistive type R, and of the inductive type L, plus ideal independent voltage and current sources E, and J. Further, we shall alow the graph of N to be separable, allowing also couplings among the components. Without a loss of generality [6], we shall consider the distribution of independent sources, called the normal distribution, such that by e- and j-shifting operations, all current sources can be put in parallel with the branches of any tree, and all voltage sources can be put in series with the links of the corresponding co-tree.

Definition: A normal tree is Bryant's tree [2,8].

Bryant's tree is constructed with all the independent voltage sources, the maximum number of capacitors, the minimum number of inductors, no independent current sources; and it is completed with resistors. Note that this is always possible, due to the properties of the normal distribution.

We shall pick a normal tree in each of the components of N. Henceforth, the collection of these normal trees shall be called the normal forest of N. We shall also classify the elements in every component using the same procedure and subscripts employed by Bryant, [2], denoting respectively by α, β, γ the capacitors, the resistors, and the inductors in the normal co-forest, and by δ, ϵ, ξ the capacitors, resistors, and inductors in the normal forest of N. Accordingly, the characteristic equations of the capacitors, resistors and inductors, (2), (1), and (3), shall be partitioned respectively as:

$$\underline{v}_k=\underline{V}_k(\underline{x}_\alpha,\underline{x}_\delta,t) \quad , \quad \underline{q}_k=\underline{Q}_k(\underline{x}_\alpha,\underline{x}_\delta,t) \quad \text{for} \quad k=\alpha,\delta. \tag{4}$$

$$\underline{v}_k=\underline{V}_k(\underline{z}_\beta,\underline{z}_\epsilon,t) \quad , \quad \underline{i}_k=\underline{I}_k(\underline{z}_\beta,\underline{z}_\epsilon,t) \quad \text{for} \quad k=\beta,\epsilon. \tag{5}$$

$$\underline{\psi}_k=\underline{\phi}_k(\underline{y}_\gamma,\underline{y}_\xi,t) \quad , \quad \underline{i}_k=\underline{I}_k(\underline{y}_\gamma,\underline{y}_\xi,t) \quad \text{for} \quad k=\gamma,\xi. \tag{6}$$

With the normal distribution, and using the properties of the normal forest, KVL, and KCL, one obtains:

$$\underline{v}_\alpha=-F_{\alpha\delta}\underline{v}_\delta+\underline{e}_\alpha \tag{7}$$

$$\underline{v}_\beta=-F_{\beta\delta}\underline{v}_\delta-F_{\beta\epsilon}\underline{v}_\epsilon+\underline{e}_\beta \tag{8}$$

$$\underline{v}_\gamma=-F_{\gamma\delta}\underline{v}_\delta-F_{\gamma\epsilon}\underline{v}_\epsilon-F_{\gamma\xi}\underline{v}_\xi+\underline{e}_\gamma \tag{9}$$

$$\underline{i}_\delta= F^T_{\gamma\delta}\underline{i}_\gamma+F^T_{\beta\delta}\underline{i}_\beta+F^T_{\alpha\delta}\underline{i}_\alpha+\underline{j}_\delta \tag{10}$$

$$\underline{i}_\epsilon= F^T_{\gamma\epsilon}\underline{i}_\gamma+F^T_{\beta\epsilon}\underline{i}_\beta+\underline{j}_\epsilon \tag{11}$$

$$\underline{i}_\xi= F^T_{\gamma\xi}\underline{i}_\gamma+\underline{j}_\xi \tag{12}$$

where $F_{\alpha\delta}$ etc., have the same meaning as in [2] , that is, the corresponding partition of the matrix F, which in turn is a partition of the fundamental circuit matrix $B=[U\vdots F]$ defined by the normal forest; moreover, the e's are the sum of the independent voltage sources around the loop defined by the respective link, and the j's are the sum of the independent current sources in the cut-set defined by the corresponding tree branch.

Defining $\underline{\psi}=\underline{\psi}_\gamma+F_{\gamma\xi}\underline{\psi}_\xi$ as the sum of the fluxes around every γ-loop; $\underline{q}=\underline{q}_\delta-F^T_{\alpha\delta}\underline{q}_\alpha$ as the sum of the charges of every δ-cut-set [6,7]; combining

(4-6) with (7-12); using the relation $\underline{v}_L(t)=d\underline{\psi}(t)/dt$, $\underline{i}_C(t)=d\underline{q}(t)/dt$; and rearranging terms we finally obtain:

$$\left.\begin{aligned} \underline{E}_\alpha(t) &= \underline{V}_\alpha(\underline{x}_\alpha,\underline{x}_\delta,t)+F_{\alpha\delta}\underline{V}_\delta(\underline{x}_\alpha,\underline{x}_\delta,t) \\ \underline{q} &= \underline{Q}_\delta(\underline{x}_\alpha,\underline{x}_\delta,t)-F^T_{\alpha\delta}\underline{Q}_\alpha(\underline{x}_\alpha,\underline{x}_\delta,t) \end{aligned}\right. \qquad \text{(C)}$$

$$\left.\begin{aligned} \underline{\psi} &= \underline{\phi}_\gamma(\underline{y}_\gamma,\underline{y}_\xi,t)+F_{\gamma\xi}\underline{\phi}_\xi(\underline{y}_\gamma,\underline{y}_\xi,t) \\ \underline{J}_\xi(t) &= \underline{I}_\xi(\underline{y}_\gamma,\underline{y}_\xi,t)-F^T_{\gamma\xi}\underline{I}\ (\underline{y}_\gamma,\underline{y}_\xi,t) \end{aligned}\right. \qquad \text{(L)}$$

$$\left.\begin{aligned} \underline{E}^*_\beta &= \underline{V}_\beta(\underline{z}_\beta,\underline{z}_\epsilon,t)+F_{\beta\epsilon}\underline{V}_\epsilon(\underline{z}_\beta,\underline{z}_\epsilon,t) \\ \underline{J}^*_\epsilon &= \underline{I}_\epsilon(\underline{z}_\beta,\underline{z}_\epsilon,t)-F^T_{\beta\epsilon}\underline{I}_\epsilon(\underline{z}_\beta,\underline{z}_\epsilon,t) \end{aligned}\right. \qquad \text{(R)}$$

and

$$d\underline{\psi}(t)/dt=-F_{\gamma\delta}\underline{V}_\delta(\underline{x}_\alpha,\underline{x}_\delta,t)-F_{\gamma\epsilon}\underline{V}_\epsilon(\underline{z}_\beta,\underline{z}_\epsilon,t)+\underline{E}_\gamma(t) \qquad (13)$$

$$d\underline{q}(t)/dt= F^T_{\gamma\delta}\underline{I}_\gamma(\underline{y}_\gamma,\underline{y}_\xi,t)+F^T_{\beta\delta}\underline{I}_\beta(\underline{z}_\beta,\underline{z}_\epsilon,t)+\underline{J}_\delta(t) \qquad (14)$$

and

$$\underline{E}^*_\beta=-F_{\beta\delta}\underline{V}_\delta(\underline{x}_\alpha,\underline{x}_\delta,t)+\underline{E}_\beta(t) \qquad (15)$$

$$\underline{J}^*_\epsilon= F^T_{\gamma\epsilon}\underline{I}_\gamma(\underline{y}_\gamma,\underline{y}_\xi,t)+\underline{J}_\epsilon(t) \qquad (16)$$

The equations (C), (L), and (R) are essentially the same as the ones obtained in [7], nevertheless, using the parametric representation allows us to write these equations in explicit form so that Palais' theorem can be applied. (see Theorem I or Appendix)

We shall represent the transformations defined by (C), (L), and (R) in vector notation as

$$\underline{s} = \underline{f}(\underline{u},t) \qquad (17)$$

where

$$\underline{s}=(\underline{E}_\alpha,\underline{q}),\quad (\underline{\psi},\underline{J}_\xi),\quad \text{or,}\quad (\underline{E}^*_\beta,\underline{J}^*_\epsilon)$$

$$\underline{u}=(\underline{x}_\alpha,\underline{x}_\delta),\quad (\underline{y}_\gamma,\underline{y}_\xi),\quad \text{or,}\quad (\underline{z}_\beta,\underline{z}_\epsilon),\ \text{respectively.}$$

b. Existence Theorem I and Corollaries.

In this section we shall prove that:

Theorem I. Given a (separable) network N, a normal forest of N, and a set of parametric representations (1-3), if the following hypothesis are satisfied:

H-1) The domain of $\underline{f}$ is the entire linear space.
H-2) $\det(\partial\underline{f}/\partial\underline{u})$ never vanishes for all $\underline{u}$ and t.
H-3) $\lim \|\underline{f}\| = \infty$ as $\|\underline{u}\| \to \infty$ for every t.
H-4) All the functions $\underline{V}_R,\underline{I}_R,\underline{V}_C,\underline{Q}_C,\underline{\phi}_L,\underline{I}_L$ are of class C^1.
H-5) All sources are regulated functions of t.

Then for any normal distribution of the independent sources:

R-1) The variables $(\underline{q},\underline{\psi})$ form a complete set [7].
R-2) A normal form exists with $(\underline{q},\underline{\psi})$ as the dependent variables of the differential equation; defined in the entire $(\underline{q},\underline{\psi})$-space.
R-3) The set $(\underline{q},\underline{\psi})$ is the state of the network taking the ind. sources

as the input and $(\underline{q},\underline{\psi})$, or any function of $(\underline{q},\underline{\psi})$, as the output. (Note that the possibility of finite escape time has not been ruled out)

Proof: By R. S. Palais' result, with t as a parameter, H-2 , H-3, and H-4 imply that the inverse of $\underline{f}$ exists and is of the class C^1. (see Appendix), i.e., we have

$$\underline{x}_k = \underline{X}_k(\underline{q},\underline{E}_\alpha(t),t) \quad \text{with } k=\alpha,\delta. \tag{18}$$

$$\underline{y}_k = \underline{Y}_k(\underline{\psi},\underline{J}_\xi(t),t) \quad \text{with } k=\gamma,\xi. \tag{19}$$

$$\underline{z}_k = \underline{Z}_k(\underline{E}^*_\beta,\underline{J}^*_\epsilon,t) \quad \text{with } k=\beta,\epsilon. \tag{20}$$

and $\underline{X}$, $\underline{Y}$, $\underline{Z}$ are of class C^1.

From (15,16) and H-4 , E^*_β and J^*_ϵ are continuously differentiable functions of $\underline{x}_\alpha,\underline{x}_\delta,\underline{y}_\gamma,\underline{y}_\xi,\underline{E}_\alpha,\underline{J}_\xi$, and t. Then by (18-2o), Z_β, Z_ϵ, are continuously differentiable functions of q, ψ, $E_\alpha, J_\xi, E_\beta, J_\epsilon$, and t. Thus, all the characteristic parameters $\underline{x},\underline{y}$, and $\underline{z}$ are uniquely determined by $\underline{q},\underline{\psi}$, and the independent sources, at every t; thus, by (1-3), all the branch characteristic variables are uniquely determined. Hence, R-1 is proved.

Substituting (18-20) via (15,16) into (13,14), we obtain the required normal form, whose right-hand side consists of functions in class $C^1$with respect to $\underline{q}$, and $\underline{\psi}$; and are regulated functions of t. Hence, R-2, and R-3 are proved.

It is noted that the set $(\underline{q},\underline{\psi})$ was obtained through a normal forest. However, the set $(\underline{q},\underline{\psi})$, complete or not, is related to any other set $(\underline{q}^*,\underline{\psi}^*)$ by a linear transformation [11], becoming in this sense independent of the choice of the normal forest. This important observation was first made by Se Jeung Ho [7].

Ranking E,C,R,L,J in this order [6], with E as the higher ranked group; we shall understand N_W (W = C,R,L) to mean the components obtained from N , when all the elements in the groups with lower rank than W have been open-circuited, and the ones with higher rank than W have been short-circuited. Clearly, all the components of N_W consist only of elements of W. For example, N_R consists of all the components of N when inductors and current sources have been open-circuited and all the voltage sources and capacitors have been short-circuited. We point out that the right-hand side of equations (C), (L), and (R) correspond to the f-circuits and f-cut-set equations of N_C, N_L, and N_R, respectively, with respect to the corresponding subsets of the normal forest.

Due to the similarity among (C), (L), and (R), an analogy can be established [6] among N_C, N_L, and N_R, thus we need only to consider the case (R). The resulting components of N_R, in general, are of three types: (i) self-loops, (ii) open-branches, (iii) components with more than one element. Further we partition the branch characteristic variables according to types (i), (ii), and (iii) : $(\underline{v}_{\beta2},\underline{i}_{\beta2}),(\underline{v}_{\epsilon2},\underline{i}_{\epsilon2})$, and $(\underline{v}_{\beta1},\underline{i}_{\beta1},\underline{v}_{\epsilon1},\underline{i}_{\epsilon1})$, and correspondigly

$$F_{\beta\epsilon} = \left[\begin{array}{c|c} F_{\beta1\epsilon1} & 0 \\ \hline 0 & 0 \end{array}\right] \tag{21}$$

We shall further assume that the characteristics can be expressed as

$$\begin{aligned} \underline{i}_{\beta1} &= \underline{I}_{\beta1}(\underline{i}_{\epsilon1},\underline{v}_{\beta1},\underline{v}_{\beta2},t) , & \underline{v}_{\epsilon1} &= \underline{V}_{\epsilon1}(\underline{i}_{\epsilon1},\underline{i}_{\epsilon2},\underline{v}_{\beta2},t) \\ \underline{i}_{\beta2} &= \underline{I}_{\beta2}(\underline{i}_{\epsilon},\underline{v}_{\beta},t) & \underline{v}_{\epsilon2} &= \underline{V}_{\epsilon2}(\underline{i}_{\epsilon},\underline{v}_{\beta},t) \end{aligned} \tag{21}$$

by imposing some conditions on the couplings.

Similarly, for the capacitive and inductive elements, we shall assume that the characteristics can be expressed as:

$$\underline{q}_{\alpha 1}=\underline{Q}_{\alpha 1}(\underline{v}_{\alpha 1},\underline{q}_{\delta 2},t)\ ;\ \underline{v}_{\delta 1}=\underline{V}_{\delta 1}(\underline{q}_{\delta 1},\underline{q}_{\delta 2},t)\ ;\ \underline{v}_{\delta 2}=\underline{V}_{\delta 2}(\underline{v}_{\alpha 1},\underline{q}_{\delta 1},\underline{q}_{\delta 2},t) \qquad (23)$$

$$\underline{\psi}_{\xi 1}=\underline{\emptyset}_{\xi 1}(\underline{i}_{\xi 1},\underline{\psi}_{\gamma 2},t)\ ;\ \underline{i}_{\gamma 1}=\underline{I}_{\gamma 1}(\underline{\psi}_{\gamma 1},\underline{\psi}_{\gamma 2},t)\ ;\ \underline{i}_{\gamma 2}=\underline{I}\ _{2}(\underline{i}_{\xi 1},\underline{\psi}_{\gamma 2},\underline{\psi}_{\gamma 1},t) \qquad (24)$$

respectively. Note that there are no self loops in N_C, and no open-branches in N_L. We can state this as:

Corollary I. Sufficient conditions to satisfy H-2 of Theorem I are:

i). All branches of the normal forest are i- or q-controlled, and all branches of the normal co-forest are v- or ψ-controlled.

ii). No couplings are allowed between the elements of the normal forest and the elements of the corresponding co-forest; excluding all those elements that correspond to self-loops and open-branches, of N_C, N_R, N_L, which can be coupled to all other elements.

iii) The Jacobians $(\partial\underline{V}_{\epsilon 1}/\partial\underline{i}_{\epsilon 1})$, $(\partial\underline{I}_{\beta 1}/\partial\underline{v}_{\beta 1})$, $(\partial\underline{V}_{\delta 1}/\partial\underline{q}_{\delta 1})$, $(\partial\underline{Q}_{\alpha 1}/\partial\underline{v}_{\alpha 1})$ $(\partial\underline{\emptyset}_{\xi 1}/\partial\underline{i}_{\xi 1})$, and $(\partial\underline{I}_{\gamma 1}/\partial\underline{\psi}_{\gamma 1})$ are positive definite (symmetric) with respect to all of their arguments and for all t.

Consequently, if H-1, H-3, H-4, and H-5 are satisfied, then, for any normal distribution of sources, R-1, R-2, and R-3 follow.

Proof. Conditions (i) and (ii) imply that the characteristic equations have the form (22), (23), and (24). Due to the similarity among the equations (C), (L), and (R), it is sufficient to show that H-2 is satisfied by equations (R), which in this case gives

$$\underline{E}^*_{\beta 2}=\underline{v}_{\beta 2}$$

$$\underline{J}^*_{\epsilon 2}=\underline{i}_{\epsilon 2}$$

$$\underline{E}^*_{\beta 1}=\underline{v}_{\beta 1}+F_{\beta 1\epsilon 1}\underline{V}_{\epsilon 1}(\underline{i}_{\epsilon 1},\underline{i}_{\epsilon 2},\underline{v}_{\beta 2},t)$$

$$\underline{J}^*_{\epsilon 1}=\underline{i}_{\epsilon 1}-F^T_{\beta 1\epsilon 1}\underline{I}_{\beta 1}(\underline{i}_{\epsilon 1},\underline{v}_{\beta 1},\underline{v}_{\beta 2},t)$$

Let J be the Jacobian of the right-hand side of these equations and $J_{\epsilon 1\epsilon 1}=\partial\underline{V}_{\epsilon 1}/\partial\underline{i}_{\epsilon 1}$, $J_{\beta 1\beta 1}=\partial\underline{I}_{\beta 1}/\partial\underline{v}_{\beta 1}$. It is easy to see that

$$\det J = \det\left[\begin{array}{c|c} I & F_{\beta 1\epsilon 1}J_{\epsilon 1\epsilon 1} \\ \hline -F^T_{\beta 1\epsilon 1}J_{\beta 1\beta 1} & I \end{array}\right]$$

and by Lemma 17 of [2] , we have $\det J = \det A\, J_{\epsilon 1\epsilon 1}$, where

$$A = \left[\begin{array}{c|c} F^T_{\beta 1\epsilon 1} & I \end{array}\right]\left[\begin{array}{c|c} J_{\beta 1\beta 1} & 0 \\ \hline 0 & J^{-1}_{\epsilon 1\epsilon 1} \end{array}\right]\left[\begin{array}{c} F_{\beta 1\epsilon 1} \\ \hline I \end{array}\right]$$

condition (iii) implies that $\left[\begin{array}{c|c} J_{\beta 1\beta 1} & 0 \\ \hline 0 & J^{-1}_{\epsilon 1\epsilon 1} \end{array}\right]$ is positive definite, hence A is positive definite. Since $J_{\epsilon 1\epsilon 1}$ is also positive definite, det J never vanishes, i.e., H-2 is proved.

Up to this point we have only considered the existence of the normal form. We shall next consider a simple network for which the normal form can always be explicitly expressed in closed form in terms of the characteristic equations and the network topology i.e., in terms of the F's-matrices.

Definition. A (separable network shall be called elementary if the choice of the normal forest is unique.

An equivalent statement in terms of the network topology is:
$F_{\alpha\delta}=0$, $F_{\gamma\xi}=0$, $F_{\beta\epsilon}=0$.

Note that an elementary network is simpler than a complete network [4].

Corollary II. Let N be elementary. Moreover, let the following hypothesis be satisfied:

i). All the branches of the normal forest are i- or q-controlled, and all the elements of the co-forest are v- or ψ-controlled.
ii). All the characteristic functions are in C^1.
iii). All the sources are regulated functions of t.

Then R-1, R-2, and R-3 of Theorem I follow, and further the explicit form of the normal form is given by:

$$\dot{\underline{q}} = F^T_{\gamma\delta}\underline{I}_\gamma(\underline{\psi},t)+F_{\beta\delta}\underline{I}_\beta(\underline{v}_\beta,\underline{i}_\epsilon,t)+\underline{J}_\delta(t)$$

$$\dot{\underline{\psi}} =-F_{\gamma\delta}\underline{V}_\delta(q,t)-F_{\gamma\epsilon}V_\epsilon(\underline{v}_\beta,\underline{i}_\epsilon,t)+\underline{E}_\gamma(t)$$

where

$$\underline{v}_\beta=-F_{\beta\delta}\underline{V}_\delta(\underline{q},t)+\underline{E}_\beta(t)$$

$$\underline{i}_\epsilon= F^T_{\gamma\epsilon}\underline{I}_\gamma(\underline{\psi},t)+J_\epsilon(t).$$

Proof. Condition (i) and definition of elementary network reduces the equations (C), (L), and (R) to $\underline{q} = \underline{q}_\delta$, $\underline{\psi} = \underline{\psi}_\gamma$, $\underline{E}^*_\beta = \underline{v}_\beta$, $\underline{J}^*_\epsilon = \underline{i}_\epsilon$. Consequently, H-1, H-2, and H-3 of Theorem I are trivially satisfied. Conditions (ii) and (iii) are the same as H-4, and H-5. The explicit normal form is obtained following the procedure used in the proof of Theorem I.

Similar results are given in [7,8].

Example 1. Let us consider the circuit in Figure 1, where the characteristics of the transistor are given by $i_c=I_c(i_b,v_{ce},t)$ and $v_{be}=V_{be}(i_b,v_{ce},t)$, the non-linear coupled transformer by $i_3=I_3(\psi_3,\psi_4, t)$ $i_4=I_4(\psi_3,\psi_4,t)$, the nonlinear coupled capacitors by $v_1=V_1(q_1,q_2,t)$ and $v_2=V_2(q_1,q_2,t)$, and the independent sources by E(t) and J(t). All elements are considered to be time-varying. This is an elementary network so the choice of the normal tree is unique.

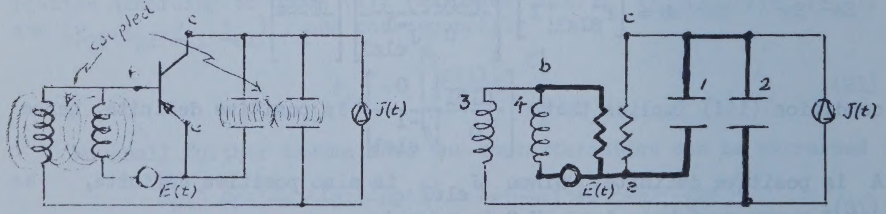


Figure 1.

Taking the cut-set equation defined by capacitors 1 and 2, and the loop equation defined by inductors 3 and 4, the differential equations are obtained as:

$$\dot{q}_1=I_c[I_3(\psi_3,\psi_4,t)+I_4(\psi_3,\psi_4,t);V_1(q_1,q_2,t),t]+I_3(\psi_3,\psi_4,t)$$

$$\dot{q}_2=I_3(\psi_3,\psi_4,t)-J(t)$$

$$\dot{\psi}_3=V_1(q_1,q_2,t)-V_2(q_1,q_2,t)+V_{be}[I_3(\psi_3,\psi_4,t)+I_4(\psi_3\psi_4,t);V_1(q_1,q_2,t)]$$

$$\dot{\psi}_4=V_{be}[I_3(\psi_3,\psi_4,t)+I_4(\psi_3,\psi_4,t);V_1(q_1,q_2,t)]+E(t)$$

which is the explicit form.

c. Existence Theorem II

The writing of the normal form in the cut-set charge $\underline{q}$, and in the loop flux $\underline{\psi}$, possesses a physical meaning which makes this approach sensible. However, there are some situations in which the normal form of the differential equations can not be written in terms of $(\underline{q},\underline{\psi})$, but can still be written in terms of the characteristic parameters.

By following a similar proof as the one given for Theorem I, we obtain the following result:

Theorem II. Given a (separable) network N, a normal forest of N and a set of parametric representations (1-3), if the following hypothesis are satisfied:

h-1. N_C contains only open-branches, and N_L contains only self-loops,

h-2. Equations (R) satisfy H-1, H-2, and H-3 of Theorem I,

h-3. H-4, and H-5 are satisfied,

h-4. Equations (C), and (L) satisfy H-1, and H-2,

Then for any normal distribution of independent sources, R-1, R-2, and R-3 are satisfied, in which we replace $(\underline{q},\underline{\psi})$ by $(\underline{x}_\delta,\underline{y}_\gamma)$.

It is implied by h-1 and h-4 that

$$\dot{\underline{x}}_\delta=(\partial\underline{Q}/\partial\underline{x}_\delta)^{-1}[\dot{\underline{q}} - (\partial\underline{Q}/\partial t)]$$

$$\dot{\underline{y}}_\gamma=(\partial\underline{\phi}/\partial\underline{y}_\gamma)^{-1}[\dot{\underline{\psi}} - (\partial\underline{\phi}/\partial t)]$$

the rest of the proof follows the proof of Theorem I. As a remark we can say that h-1 can be relaxed if we assume that we can obtain the necessary number of functions to eliminate x_α, and y_ξ from the rest of the equations.

It is not our purpose to elaborate in this alternate procedure but rather show its meaning by considering a simple case.

Example 2. Consider a pair of coupled inductors, each one of them loaded with a linear one-ohm resistor across their terminals. Let the characteristics of the inductors be given by

$$\psi_1=e^{-y_1}\sin y_2, \quad i_1=I_1(y_1,y_2)$$

and

$$\psi_2=e^{-y_1}\cos y_2, \quad i_2=I_2(y_1,y_2).$$

Immediately we see that the form of equations (13,14) is

$$\dot{\psi}_1=-I_1(y_1,y_2), \quad \dot{\psi}_2=-I_2(y_1,y_2).$$

From Theorem I. we observe that H-3 is not satisfied since

$$\lim \|\underline{f}\| = e^{-2y_1} = 0 \quad \text{as} \quad y_1 \to \infty,$$

but h-4 of Theorem II is satisfied since $(\partial \underline{\phi} / \partial \underline{y}_\gamma) \neq 0$ for all $\underline{y}_\gamma$. Therefore we can write the normal form in the characteristic parameters y_1, and y_2 as

$$\dot{y}_1 = e^{y_1}(\sin y_2) I_1(y_1, y_2) - e^{y_1}(\cos y_2) I_2(y_1, y_2)$$

$$\dot{y}_2 = e^{y_1}(\cos y_2) I_1(y_1, y_2) + e^{y_1}(\sin y_2) I_2(y_1, y_2)$$

which is established by Theorem II but not by Theorem I.

APPENDIX

The following two Corollaries are a minor revision of corollary 4.3 in 12 .

Corollary 1: Let $f_1, \ldots, f_n$ be n real functions of n real variables.

The necessary and sufficient conditions that the mapping $\underline{f}$: $R^n \to R^n$ defined by $\underline{f}(\underline{x}) = (f_1(\underline{x}), \ldots, f_n(\underline{x}))$ be a C^k diffeomorphism of R^n onto itself are:

1. each f_i is of class C^k.
2. $\det(\partial f_i / \partial x_j)$ never vanishes.
3. $\lim \|\underline{f}(\underline{x})\| = \infty$ as $\|\underline{x}\| \to \infty$.

Corollary 2: Let $f_1, \ldots, f_n$ be n real functions of n real variables.

The necessary and sufficient conditions that the mapping $\underline{f}$: $R^n \to R^n$ defined by $\underline{f}(\underline{x}) = (f_1(\underline{x}), \ldots, f_n(\underline{x}))$ be a homeomorphism of of R^n onto itself are:

1. $\underline{f}$ is local homeomorphism.
2. $\lim \|\underline{f}(\underline{x})\| = \infty$ as $\|\underline{x}\| \to \infty$.

ACKNOWLEDGMENT

The authors wish to express their appreciation to Pravin P.Varaiya, University of California, Berkeley, for his valuable suggestions and discussion. This research has been sponsored by N.S.F, GK-91.

REFERENCES

[1] Duffin, R.J.,"Nonlinear Networks"(I,II,IIa) Bull. Amer. Math.Soc. Vols.52,53,54,pp.836-838,963-971,119-127, 1946,1947,1948.

[2] Bryant, P.R.,"The Order of Complexity of Electrical Networks", Proc.IEE.,106c,pp.174-188, 1959.

[3] Liu,R.W. and Auth,L.V.,"Qualitative Synthesis of Nonlinear Networks", Proc.First Anual Allerton Conference on Circuits and Systems Theory.,pp330-343., Nov.,1963.

[4] Brayton,R.K., and Moser,J.K.,"A Theory of Nonlinear Networks" (I-II), Q.of Appl.Math.,Vol.22, pp.1-33,31-103, April and June 1964.

[5] Auth,L.V., and Liu,R.W.,"On the Existence of the Normal Form for Nonlinear Networks at Large with Controlled Sources", Proc. of the Second Allerton Conference.onCircuits and Systems Theory., pp.288-292,Sept.1964.

[6] Desoer, C.A. and Katzenelson,J.K.,"Nonlinear RLC Networks" BSTJ, Jan.,1965.

[7] Stern,T.E.,"On the Equations of Nonlinear Networks" IEEE Trans. on Circuit Theory.,to appear.

[8] Chua,L.O., and Rohrer,R.A.,"On the Dynamical Equations of a Class of Nonlinear RLC Networks" IEEE Trans.on Circuit Theory ,to appear.

[9] Liu,R.W.,"The Self-Sustained and the Forced Oscillations of Time-Invariant Nonlinear Networks at Large", ICMCI, Tokyo, Sept.1964.

[10] Kuh,E.S.,and Rohrer,R.A.,"The State Variable Approach to Network Analysis", Proc.IEEE, Vol.53, No7, July 1965.

[11] Seshu,S.,and Reed,M.B., Linear Graphs and Electrical Networks, Addison Wesley, 1961.

[12] Palais,R.S.,"Natural Operations on Differential Forms",Trans.Amer. Math. Soc.,Vol.92,pp.125-141, 1959.

A METHOD FOR STUDYING THE BEHAVIOR OF NONLINEAR SYSTEMS NEAR AND AT INFINITY OF THE PHASE PLANE

WAYNE H. CHEN
Department of Electrical Engineering
University of Florida
Gainesville, Florida

ABSTRACT

Poincare's projective transformation that maps the (x,y) plane unto a unit-sphere $\mathcal{U}$ is a known tool for the study of the system behavior for $dx/dt = P(x,y)$, $dy/dt = Q(x,y)$ near and at infinity of the phase plane. However, existing procedure is applicable only to singular points at infinity in the fixed directions along the x and y axes. Introducing auxiliary transformations that rotate the unit-sphere $\mathcal{U}$ and project the singular points on the equator unto its respective tangential planes, this paper presents a general procedure for investigating system behavior at infinity. This procedure is formalized with simple steps and requires no knowledge of the transformations for its application.

I. INTRODUCTION

The phase space or state approach has in recent years become increasingly important as a concept and tool in the analysis of linear and nonlinear circuits and systems.(1),(2),(3) A phase plane is a 2-dimensional phase space providing a pictorial description of the behavior of the system,

$$\frac{dx}{dt} = P(x,y), \quad \frac{dy}{dt} = Q(x,y) \tag{1}$$

where $P(x,y)$ and $Q(x,y)$ are polynomials in x and y with p and q as the respective highest degrees of their terms. The system in (1) is called normal if $p = q$, and for example,

$$\frac{dx}{dt} = x-2xy + x^2, \quad \frac{dy}{dt} = 2y-2xy + 2y^2 \tag{2}$$

is a normal system for $p = q = 2$. The system in (1) is called nonnormal if $p \neq q$, and for example,

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = -2x + 3y + x^2 \tag{3}$$

is a nonnormal system for $p = 1$ and $q = 2$.

The so-called "qualitative" approach suggests that we determine the approximate system behavior for (1) in the "finite" phase plane by 1) completely characterizing each singular point with respect to its type, asymptotes, approximate trajectory pattern and parametrization;

and then 2) interconnecting the trajectory patterns of these singular points, obtaining a finite phase portrait as exemplified in Figure 1. Simple procedures are available for characterizing elementary singular points$^{(4)}$ and non-elementary singular points.$^{(5)}$ These procedures are, however, inapplicable in characterizing singular points at infinity in the phase plane.

After a preliminary discussion on transformations in Section II of this paper, a method made possible by these transformations is introduced in Section III for use in investigating the singular points at infinity of the phase plane and their associated trajectory patterns for the normal systems. The investigation of nonnormal systems will be reported in a future paper. The trajectory patterns near the singular points at infinity are the asymptotic or near-asymptotic behavior of the system when the amplitude of the system motion becomes extremely large.

II. THE TRANSFORMATIONS FOR THE STUDY OF SYSTEM BEHAVIOR NEAR AND AT INFINITY

A. The Π_1 - $\not{N}_2$ Transformation

A projective transformation T_1 due to Poincare$^{(6)}$ and described by $x/1 = \tilde{x}/\tilde{z}$, $y/1 = \tilde{y}/\tilde{z}$ and $\tilde{x}^2 + \tilde{y}^2 + \tilde{z}^2 = 1$ first maps the phase plane, i.e., the Π_1 or (x,y) plane, unto the surface of a unit sphere $\not{N}_{\utilde{2}}$ in Figure 2, and the points (E) at infinity in Π_1 into the equator E of $\not{N}_2$. It transforms the non-parametrized version of (1) into

$$A(\tilde{x},\tilde{y},\tilde{z})\, d\tilde{x} + B(\tilde{x},\tilde{y},\tilde{z})\, d\tilde{y} + C(\tilde{x},\tilde{y},\tilde{z})\, d\tilde{z} = 0, \qquad (4)$$
$$\tilde{x}^2 + \tilde{y}^2 + \tilde{z}^2 = 1,$$

where, for $p = q = n$ in normal systems, $A(\tilde{x},\tilde{y},\tilde{z}) = -\tilde{z}\overline{Q}(\tilde{x},\tilde{y},\tilde{z})$, $B(\tilde{x},\tilde{y},\tilde{z}) = \tilde{z}\overline{P}(\tilde{x},\tilde{y},\tilde{z})$, $C = (\tilde{x},\tilde{y},\tilde{z}) = \tilde{x}\overline{Q}(\tilde{x},\tilde{y},\tilde{z}) - \tilde{y}\overline{P}(\tilde{x},\tilde{y},\tilde{z})$ with the homogeneous polynomials $\overline{P}(\tilde{x},\tilde{y},\tilde{z})$ and $\overline{Q}(\tilde{x},\tilde{y},\tilde{z})$ defined as follows:

$$\overline{P}(\tilde{x},\tilde{y},\tilde{z}) \text{ or } \overline{Q}(\tilde{x},\tilde{y},\tilde{z}) = \tilde{z}^n\,[\,P(x,y) \text{ or } Q(x,y)]_{x=\tilde{x}/\tilde{z};\; y=\tilde{y}/\tilde{z}}. \qquad (5)$$

B. Positions of Singular Points on the Equator $\tilde{E}$ of $\not{N}_2$

The singular points of the system in (4) on $\not{N}_2$ must satisfy $A(\tilde{x},\tilde{y},\tilde{z}) = 0$, $B(\tilde{x},\tilde{y},\tilde{z}) = 0$, $C(\tilde{x},\tilde{y},\tilde{z}) = 0$. For the singular points on the equator $\utilde{E}$ of $\not{N}_2$ in Figure 2, we further impose $\tilde{z} = 0$, obtaining

$$\tilde{x}\overline{Q}(\tilde{x},\tilde{y},0) - \tilde{y}\overline{P}(\tilde{x},\tilde{y},0) = 0. \qquad (6)$$

By factoring (6) into $K\,[(\tilde{y}-\alpha_1\tilde{x})(\tilde{y}-\alpha_2\tilde{x})\ldots(\tilde{y}-\alpha_h\tilde{x})] \times$ [complex factors], the positions of the singular points satisfy

$$\tilde{y} = \alpha_i\tilde{x}, \quad \tilde{z}=0, \quad i=1,2,\ldots,h. \qquad (7)$$

as demonstrated in Figure 3.

C. Local Properties of Singular Points on the Equator $\tilde{E}$ of $\mathcal{N}_2$

To investigate each of these singular points, say, S_i, in Figure 3, we first 1) rotate the $\tilde{x}$- and $\tilde{y}$- axes to make S_i the $(\dot{x},\dot{y},\dot{z})=(1,0,0)$ in the new coordinate system with the transformation T_a: $\tilde{x}=\dot{x}\cos\theta_i - \dot{y}\sin\theta_i = (\dot{x}-\alpha_i\dot{y})/\sqrt{1+\alpha_i^2}$, $\tilde{y} = \dot{x}\sin\theta_i + \dot{y}\cos\theta_i = (\alpha_i\dot{x}+\dot{y})/\sqrt{1+\alpha_i^2}$, $\tilde{z}=\dot{z}$; 2) project the representation on $\mathcal{N}_2$ in the neighborhood of S_i into a Π_2 plane which is tangential to $\mathcal{N}_2$ at S_i, as demonstrated in Figure 4, with the transformation T_b: $\dot{x}=1$, $\dot{y}=\hat{y}$, $\dot{z}=\hat{z}$, obtaining in Π_2 the system $\hat{Q}(\hat{y},\hat{z})\,d\hat{y} - \hat{P}(\hat{y},\hat{z})\,d\hat{z} = 0$ or its parametrized version:

$$\frac{d\hat{y}}{d\tau} = \hat{P}(\hat{y},\hat{z}), \quad \frac{d\hat{z}}{d\tau} = Q(\hat{y},\hat{z}). \tag{8}$$

Combining transformations T_a and T_b, we have

$$\tilde{x} = (1-\alpha_i\hat{y})/\sqrt{1+\alpha_i^2}, \quad \tilde{y} = (\alpha_i+\hat{y})/\sqrt{1+\alpha_i^2}, \quad \tilde{z} = \hat{z} \tag{9}$$

and

$$\begin{aligned} \hat{P}(\hat{y},\hat{z}) &= -\left[\tilde{x}\overline{Q}(\tilde{x},\tilde{y},\tilde{z}) - \tilde{y}\overline{P}(\tilde{x},\tilde{y},\tilde{z})\right]_{\tilde{x},\tilde{y},\tilde{z}\text{ in }(9)} \\ \hat{Q}(\hat{y},\hat{z}) &= (\tilde{z}/\sqrt{1+\alpha_i^2})\left[\alpha_i\overline{Q}(\tilde{x},\tilde{y},\tilde{z}) + \overline{P}(\tilde{x},\tilde{y},\tilde{z})\right]_{\tilde{x},\tilde{y},\tilde{z}\text{ in }(9)}. \end{aligned} \tag{10}$$

By rearranging the terms of (10), the system in (8) has one of these two forms:

$$\frac{d\hat{y}}{d\tau} = A\hat{y} + 0 + \left[\begin{matrix}\text{non-linear}\\ \text{terms}\end{matrix}\right], \quad \frac{d\hat{z}}{d\tau} = 0 + D\hat{z} + \left[\begin{matrix}\text{non-linear}\\ \text{terms}\end{matrix}\right], \tag{11}$$

$$\text{or}\quad \frac{d\hat{y}}{d\tau} = A\hat{y} + B\hat{z} + \left[\begin{matrix}\text{non-linear}\\ \text{terms}\end{matrix}\right], \quad \frac{d\hat{z}}{d\tau} = 0 + D\hat{z} + \left[\begin{matrix}\text{non-linear}\\ \text{terms}\end{matrix}\right]. \tag{12}$$

Some basic information about the singular point S_i at the origin associated with (11) or (12) is readily obtainable:(2),(4) and given as follows: 1) S_i is a node if A and D are of the same sign; and is a saddle point, if of different signs. 2) S_i has

$$\begin{aligned} &L_1 \text{ for both (11) and (12): } \hat{z} = 0 \text{ (i.e., the } \hat{y}\text{-axis)} \\ &L_2 \text{ for (11): } \hat{y} = 0, \qquad L_2 \text{ for (12): } (A-D)\hat{y} + B\hat{z} = 0, \end{aligned} \tag{13}$$

as its asymptotic tangents. 3) When S_i is a node, L_1 in (13) is the major asymptotic tangent if $|D| > |A|$, or L_2 is the major asymptotic tangent if $|A| > |D|$.

Knowing the trajectory pattern in the neighborhood of the singular point S_i in Π_2, we know the trajectory pattern in the neighborhood of the singular point S_i on the equator $\tilde{E}$ of $\mathcal{N}_2$; and through the transformation T_1 in Figure 2, we also know the trajectory pattern near the singular point S_i at infinity of the Π_1 or (x,y) plane. We may repeat this investigation for all singular points at infinity of the Π_1 of (x,y) plane.

In the above discussion, all singular points are assumed to be elementary. If some of them are non-elementary, modified procedures$^{(5)}$ must be followed.

III. THE PROCEDURE FOR STUDYING THE COMPLETE SYSTEM BEHAVIOR

We are now interested in the complete system behavior of (1), as described by the "complete phase portrait" of the system. To obtain this complete phase portrait, we first 1) obtain the finite phase portrait with the procedure as described in Section I; 2) obtain the trajectory patterns near the singular points at infinity with the procedure given below and based upon the transformations described in Section II; and then 3) interconnect them to obtain the complete phase portrait. Illustrations are provided in Section IV.

A. Step 1: Determine the Locations of the Singular Points at Infinity

We first 1) draw a large circle E (of course, not to scale) to represent the points at infinity of the phase plane; 2) obtain the real linear factors

$$y - \alpha_i x = 0 \quad \text{or} \quad y = \alpha_i x, \qquad i = 1,2,\ldots,h, \tag{14}$$

of

$$xQ_n(x,y) - yP_n(x,y) = 0, \tag{15}$$

where $P_n(x,y)$ and $Q_n(x,y)$ are the highest-degree terms of $P(x,y)$ and $Q(x,y)$ in (1); and 3) draw straight lines with (14) to intersect E at S_i and S_i' as exemplified in Figure 6, where S_i and S_i' are the singular points at infinity.

Equations (6) and (15) are the $\mathcal{N}_2$ and Π_1 versions of the same relation through the transformation T_1 in Figure 2; and Step 1, as given here, is therefore justified.

B. Step 2: Obtain the "Local Portraits" of the Singular Points

We first 1) obtain the homogeneous polynomials $\overline{P}(\tilde{x},\tilde{y},\tilde{z})$ and $\overline{Q}(\tilde{x},\tilde{y},\tilde{z})$ from $P(x,y)$ and $Q(x,y)$ of the given system in (1) with (5); 2) obtain $\hat{P}(\hat{y},\hat{z})$ and $\hat{Q}(\hat{y},\hat{z})$ in (8) with (10), separating their linear terms in the form of $A\hat{y} + 0$ (or $A\hat{y} + B\hat{z}$) and $0 + D\hat{z}$ from non-linear terms and obtaining the equations of the singular point S_i in the form of (11) or (12); and 3) sketch the local phase portrait of S_i in the Π_2 plane with the information given in the paragraph following (12), as exemplified in Figure 5.

Repeat this step for all singular points S_i.

C. Step 3: Transplant the "Local Portraits" to the Phase Plane

We "transplant" or "resketch" 1) the upper half of the local portrait of S_i as obtained above to represent the trajectory pattern near S_i on E at infinity and 2) the flipover version of the lower half to represent the trajectory pattern near S_i' on E at infinity. For example, the local portraits in Figure 5 are transplanted to Figure 6.

IV. ILLUSTRATIONS

A. Illustration 1

For

$$\frac{dx}{dt} = P(x,y) = x - 2xy + x^2, \quad \frac{dy}{dt} = Q(x,y) = 2y - 2xy + y^2, \qquad (16)$$

we first obtain its finite phase portrait in Figure 1.

Following Step 1 (Section III-A), we first use (15) obtaining $x(-2xy + 2y^2) - y(-2xy + x^2) = 4(y-0)\,[y - (3/4)\,x]\,(x-0) = 0$. This implies $y = \alpha_i x$ for $\alpha_1 = 0$, $\alpha_2 = 3/4$, $\alpha_3 = \infty$. Drawing straight lines $y = \alpha_i x$ to intersect the circle E at infinity, we determine singular points S_1, S_1'; S_2, S_2'; S_3, S_3' in Figure 6.

Following Step 2 (Section III-B), we first obtain $\bar{P}(\tilde{x},\tilde{y},\tilde{z}) = \tilde{x}\tilde{z} - 2\tilde{x}\tilde{y} + \tilde{x}^2$ and $\bar{Q}(\tilde{x},\tilde{y},\tilde{z}) = 2\tilde{y}\tilde{z} - 2\tilde{x}\tilde{y} + 2\tilde{y}^2$. For singular point S_1 and $\alpha_1 = 0$. We use (8), (9), and (10) obtaining

$$\frac{d\hat{y}}{d\tau} = 3\hat{y} - \hat{y}\,(\hat{z} + 4\hat{y}), \quad \frac{d\hat{z}}{d\tau} = \hat{z} + \hat{z}(\hat{z} - 4\hat{y} + 2\hat{y}\hat{z} + 2\hat{y}^2), \qquad (17)$$

with which we sketch the local portrait of S_1 in Figure 5(a). Similarly, we obtain $d\hat{y}/d\tau = -(12/5)\,\hat{y} - (12/25)\,\hat{z}$ + [non-linear terms], $d\hat{z}/d\tau = -(2/5)\hat{z}$ + [non-linear terms], and the local portrait of S_2 in Figure 5(b); and $d\hat{y}/d\tau = 4\hat{y}$ + [non-linear terms], $d\hat{z}/d\tau = 2\hat{z}$ + [non-linear terms], and the local portrait of S_3 in Figure 5(c).

Following Step 3 (Section III-C), we obtain the trajectory patterns near all the singular points at infinity for the system in (16).

Combining and interconnecting Figures 1 and 6, we obtain the complete phase portrait of the system in (16) in Figure 7.

B. Illustration 2

For

$$\frac{dx}{dt} = -6x + y + y^2, \quad \frac{dy}{dt} = y + x^2, \qquad (18)$$

we use the same procedure, first 1) obtaining its finite phase portrait in Figure 8(a); 2) determining the locations of its singular points S_1 and S_1' in Figure 9 and the local portrait of S_1 in Figure 8(b); 3) obtaining the trajectory patterns near S_1 and S_1' at infinity in Figure 9; and 4) combining and interconnecting Figures 8(a) and (9) to obtain the complete phase portrait of the system in (18) in Figure 10.

C. Illustration 3

For

$$\frac{dx}{dt} = 3x^2 - y^2, \quad \frac{dy}{dt} = -x^2 + 4xy - y^2 \tag{19}$$

we use similar general procedure, except for the finite singular point S_a which is non-elementary and requires special attention,$^{(5)}$ to obtain its complete phase portrait in Figure 11.

REFERENCES

1. A.A. Andronow and C.E. Chaikin, "Theory of Oscillations", (English language edition), Princeton University Press, Princeton, N.J., 1949.

2. Y.H. Ku, "Analysis and Control of Nonlinear Systems", The Ronald Press Co., New York, N.Y.; 1958.

3. L.A. Zadeh and C.A. Desoer, "Linear Systems Theory: The State Space Approach", McGraw-Hill Book Co., Inc., New York, N.Y.; 1963.

4. N. Minorsky, "Theoretical Aspects of Nonlinear Oscillations", IRE Trans. on Circuit Theory, Vol. CT-7, pp. 368-381; December, 1960.

5. W.H. Chen, "Characterization of Non-elementary Singular Points of Nonlinear Systems in the Phase Plane", Proceedings of the Second Annual Allerton Conference on Circuit and System Theory (September 28-30, University of Illinois), pp. 250-272; 1964.

6. H. Poincare, "Sur les courbes definies par une equation differentielle", Journal de Mathematiques, Vol. 8, pp. 251-296; 1882. Also, "Oeuvres", Vol. 1, Gauthier-Villars, Paris; 1892.

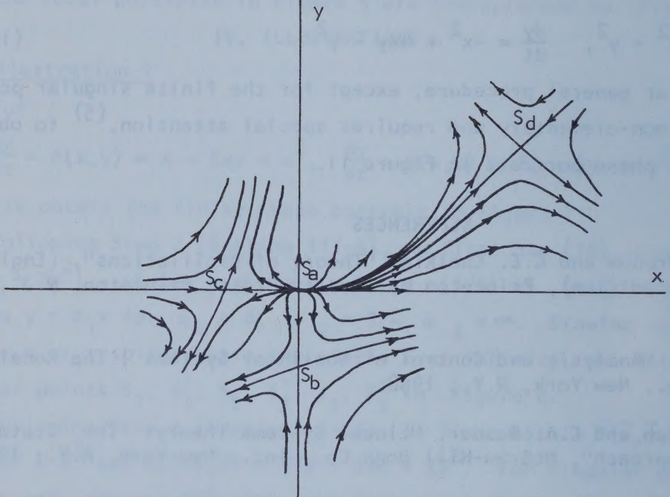
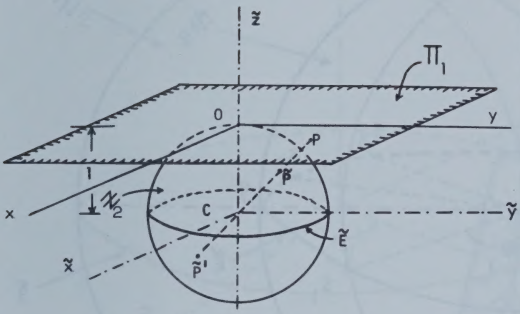

y
x
S_a
S_b
S_c
S_d

Fig. 1 - A Phase Portrait in the Finite Phase Plane


$\tilde{z}$
Π_1
0
y
P
$\tilde{P}$
1
$\mathcal{H}_2$
x
C
$\tilde{y}$
$\tilde{E}$
$\tilde{x}$
$\tilde{P}'$

Π_1	$\mathcal{H}_2$
P at (x,y)	P at $(\tilde{x},\tilde{y},\tilde{z})$ & $\tilde{P}'$ at $(-\tilde{x},-\tilde{y},-\tilde{z})$
O at $(0,0)$	O at $(0,0,1)$
Points at infinity (E)	Points on Equator $\tilde{E}$

Fig. 2 - Transformation T_1

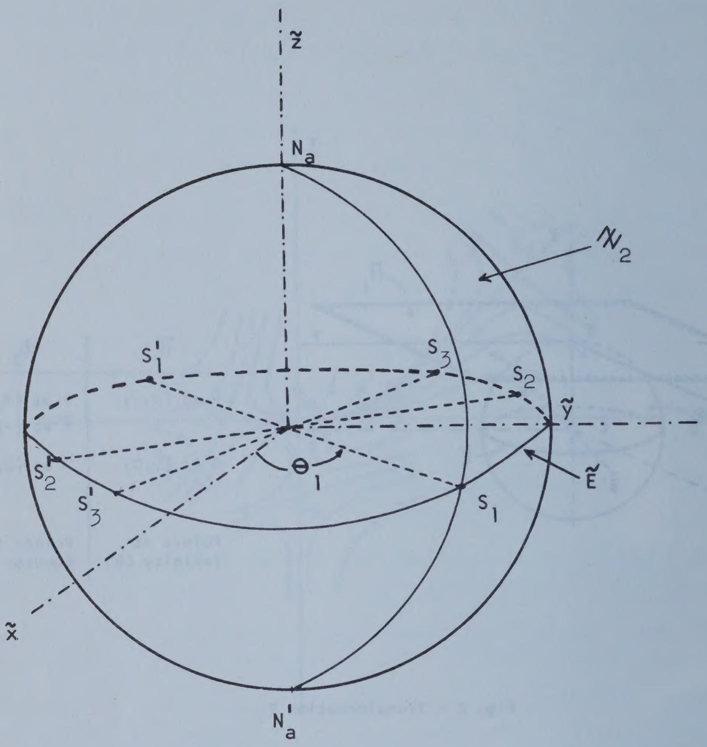

$\tilde{z}$
N_a
$\mathscr{N}_2$
S_1'
S_3
S_2
$\tilde{y}$
S_2'
θ_1
$\tilde{E}$
S_3'
S_1
$\tilde{x}$
N_a'

Fig. 3 - Positions of Singular Points on the Equator $\tilde{E}$ of $\mathscr{N}_2$

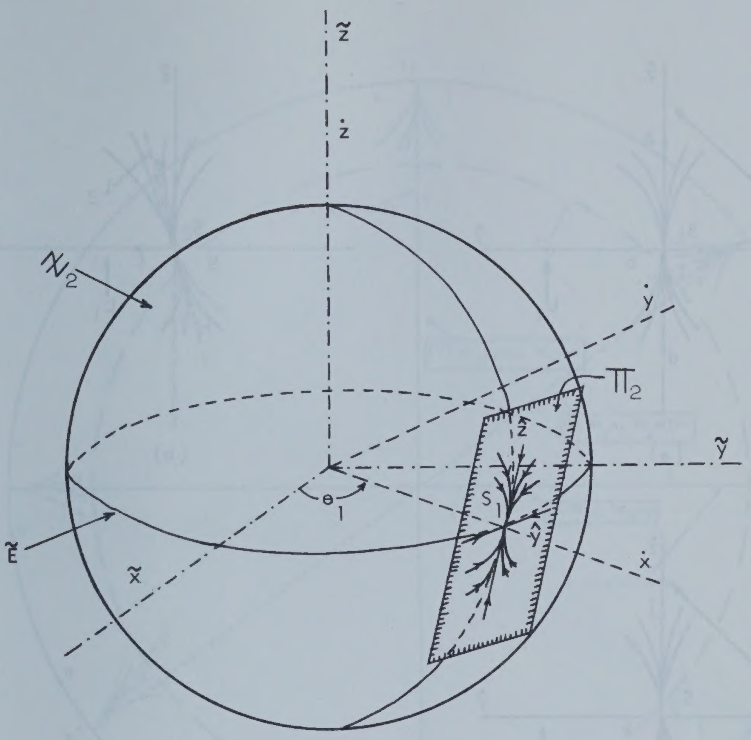
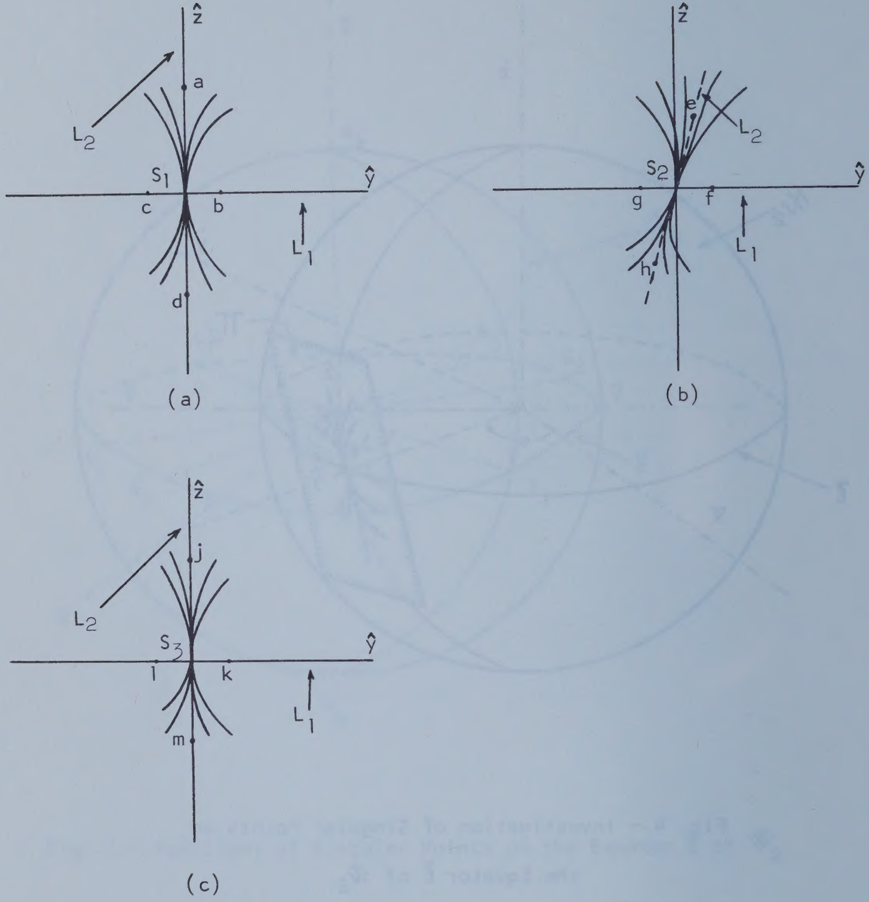

$\tilde{z}$
$\dot{z}$
$\mathcal{N}_2$
$\dot{y}$
Π_2
$\hat{z}$
$\tilde{y}$
S_1
θ_1
$\hat{y}$
$\tilde{E}$
$\dot{x}$
$\tilde{x}$

Fig. 4 - Investigation of Singular Points on the Equator $\tilde{E}$ of $\mathcal{N}_2$



$\hat{z}$
a
L_2
S_1
$\hat{y}$
c
b
L_1
d
(a)
$\hat{z}$
e
L_2
S_2
$\hat{y}$
g
f
L_1
h
(b)
$\hat{z}$
j
L_2
S_3
$\hat{y}$
l
k
L_1
m
(c)

Fig. 5 - Singular Points of the System in (16)

at the Intersections of $y=\alpha_i x$ and E at infinity.

(a) S_1 for $\alpha_1=0$. (b) S_2 for $\alpha_2=3/4$. (c) S_3 for $\alpha_3=\infty$.

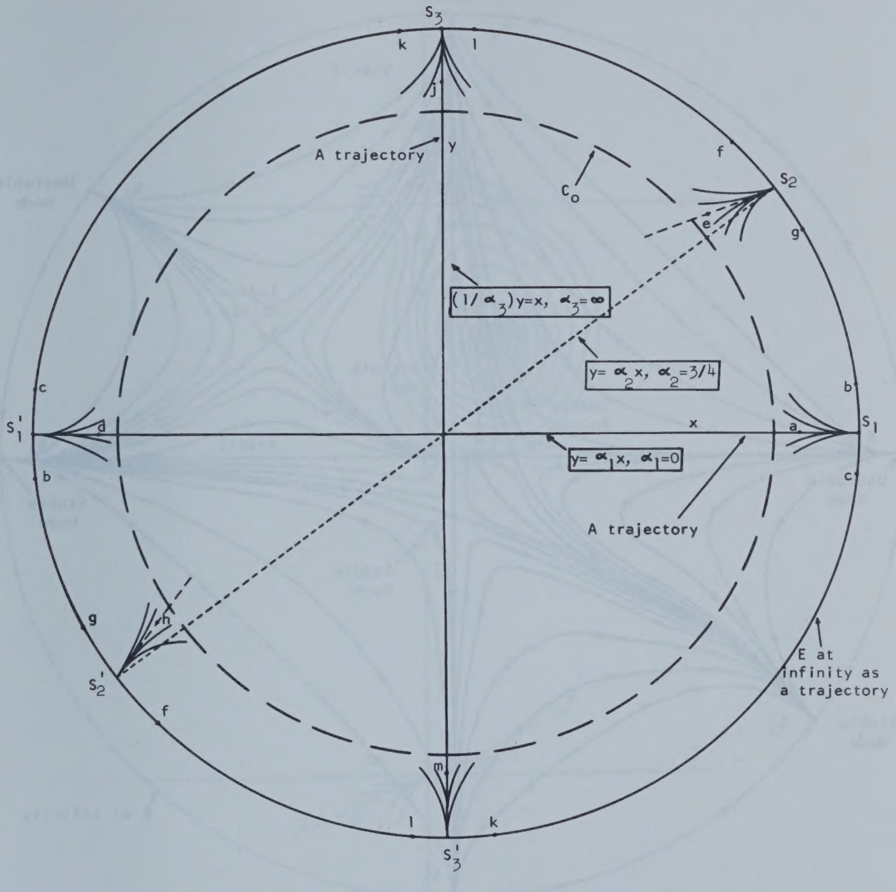

S_3
k
l
j
A trajectory
y
f
S_2
C_o
e
g
$(1/\alpha_3)y=x,\ \alpha_3=\infty$
$y=\alpha_2 x,\ \alpha_2=3/4$
c
b
S_1'
d
x
a
S_1
$y=\alpha_1 x,\ \alpha_1=0$
b
c
A trajectory
g
h
E at infinity as a trajectory
S_2'
f
m
l
k
S_3'

Fig. 6 - Trajectory Patterns Near the Singular Points S_1, S_2, S_3, S_1', S_2', S_3' at Infinity for the System in (16) as Transplanted From Fig. 5.

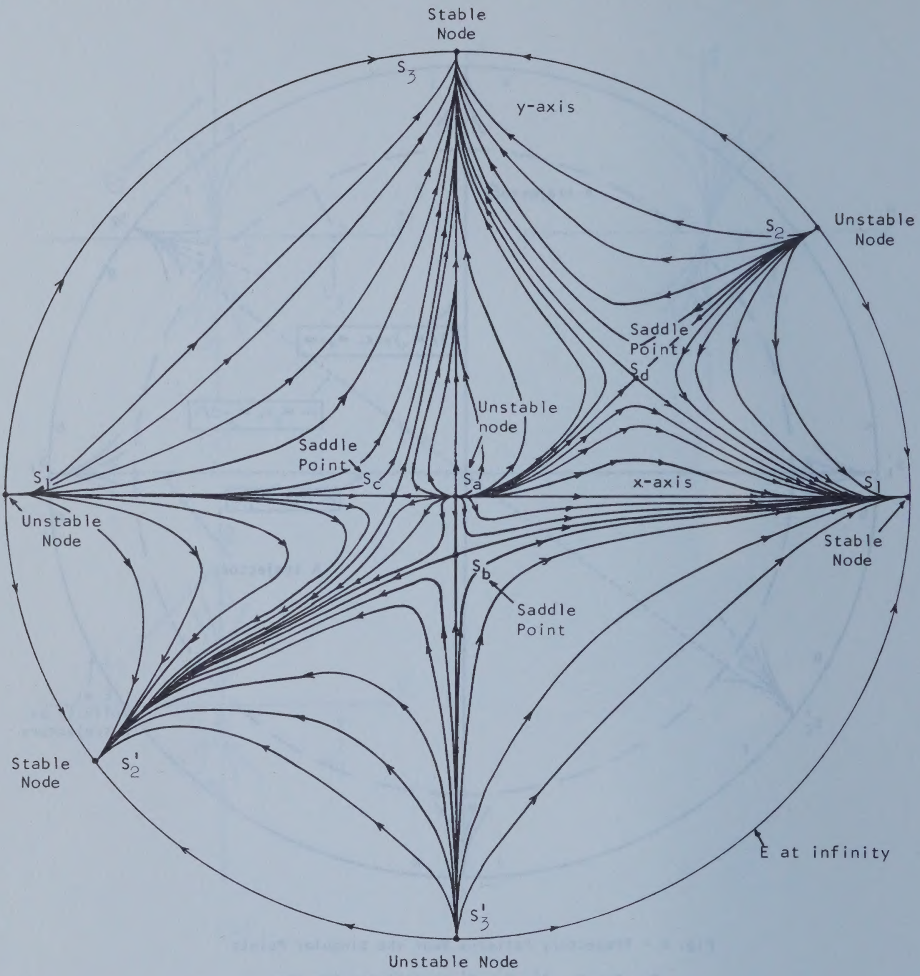

Stable
Node
S_3
y-axis
S_2
Unstable
Node
Saddle
Point
S_d
Unstable
node
Saddle
Point
S_c
S_a
x-axis
S_1'
S_1
Unstable
Node
Stable
Node
S_b
Saddle
Point
Stable
Node
S_2'
E at infinity
S_3'
Unstable Node

Fig. 7 - The Complete Phase Portrait of the System in (16).

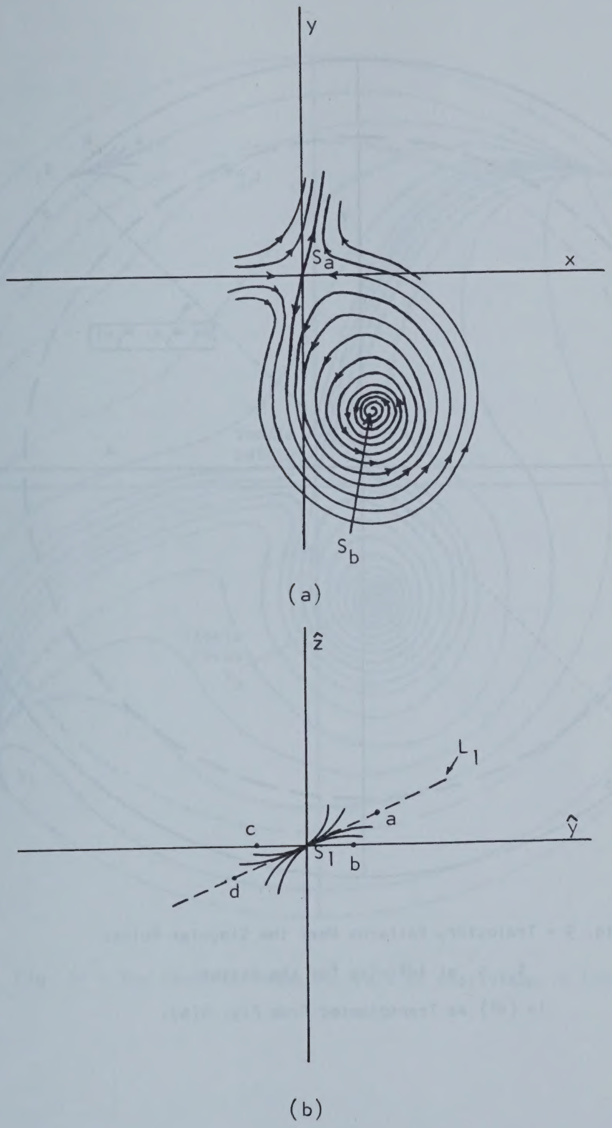

y
x
S_a
S_b
(a)
$\hat{z}$
$\hat{y}$
L_1
a
b
c
d
S_1
(b)

Fig. 8 - Elementary Singular Points of the System in (18).

(a) S_a at (0,0) and S_b at (2,-4) in the Finite Phase Plane.

(b) S_1 at the intersection of $y = \alpha_1 x$ and E at infinity for $\alpha_1 = 1$.

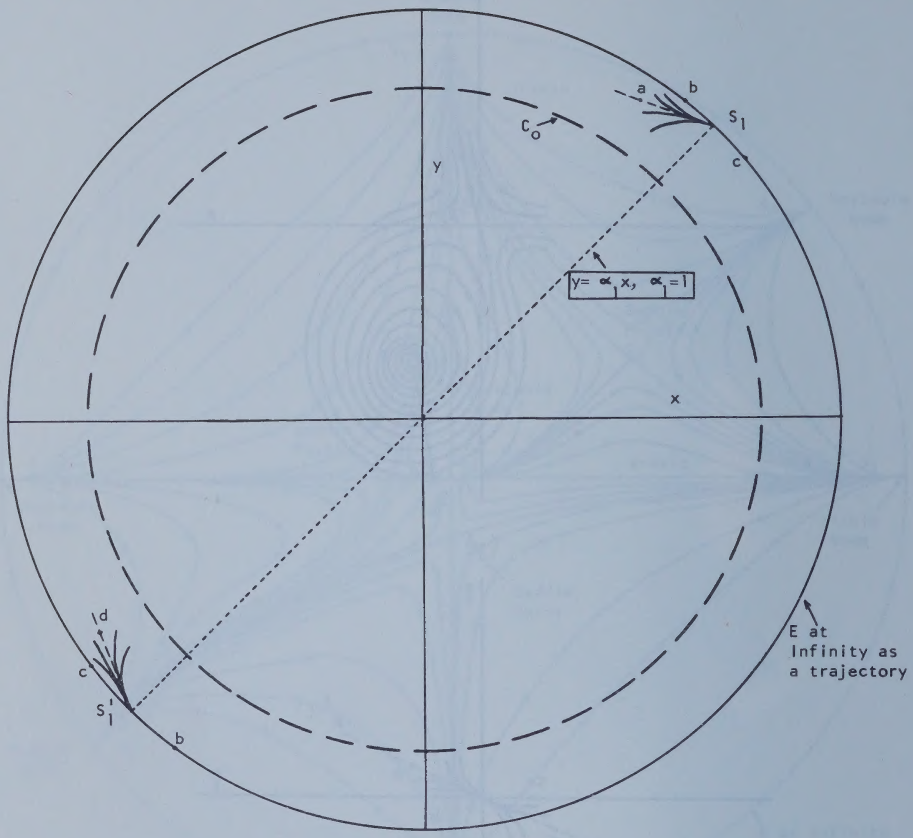

a
b
S_1
C_o
y
c
$y = \alpha_1 x, \ \alpha_1 = 1$
x
d
c
S_1'
b
E at
Infinity as
a trajectory

Fig. 9 - Trajectory Patterns Near the Singular Points S_1, S_1' at Infinity for the System in (18) as Transplanted From Fig. 8(b).

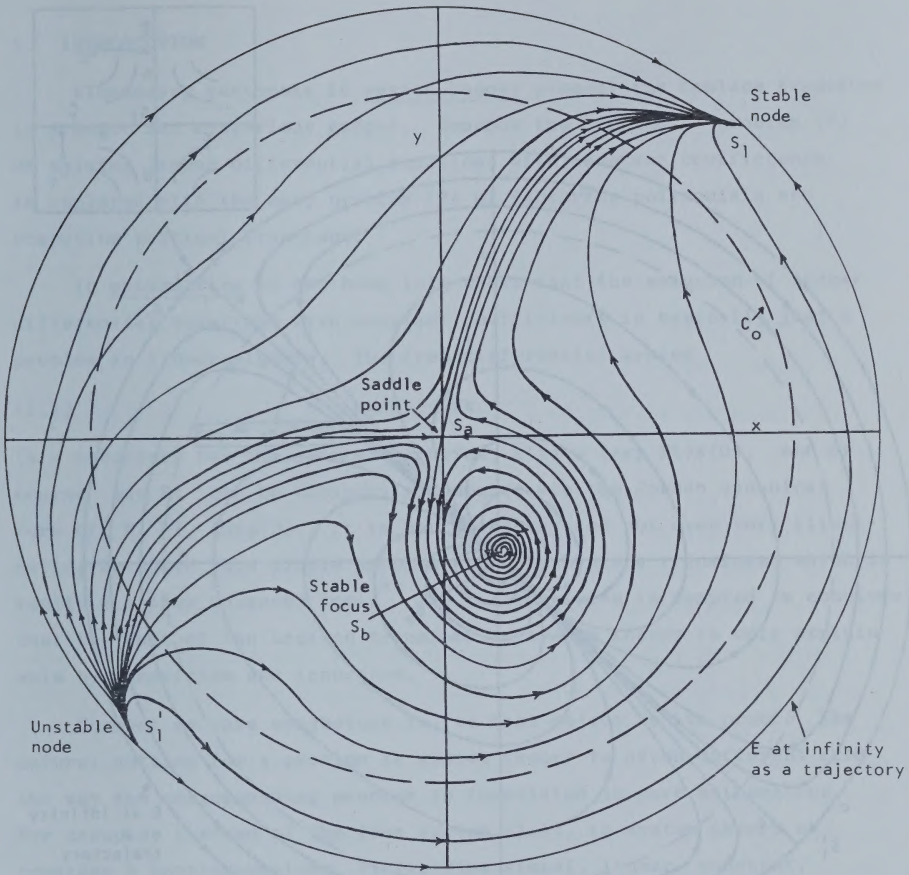
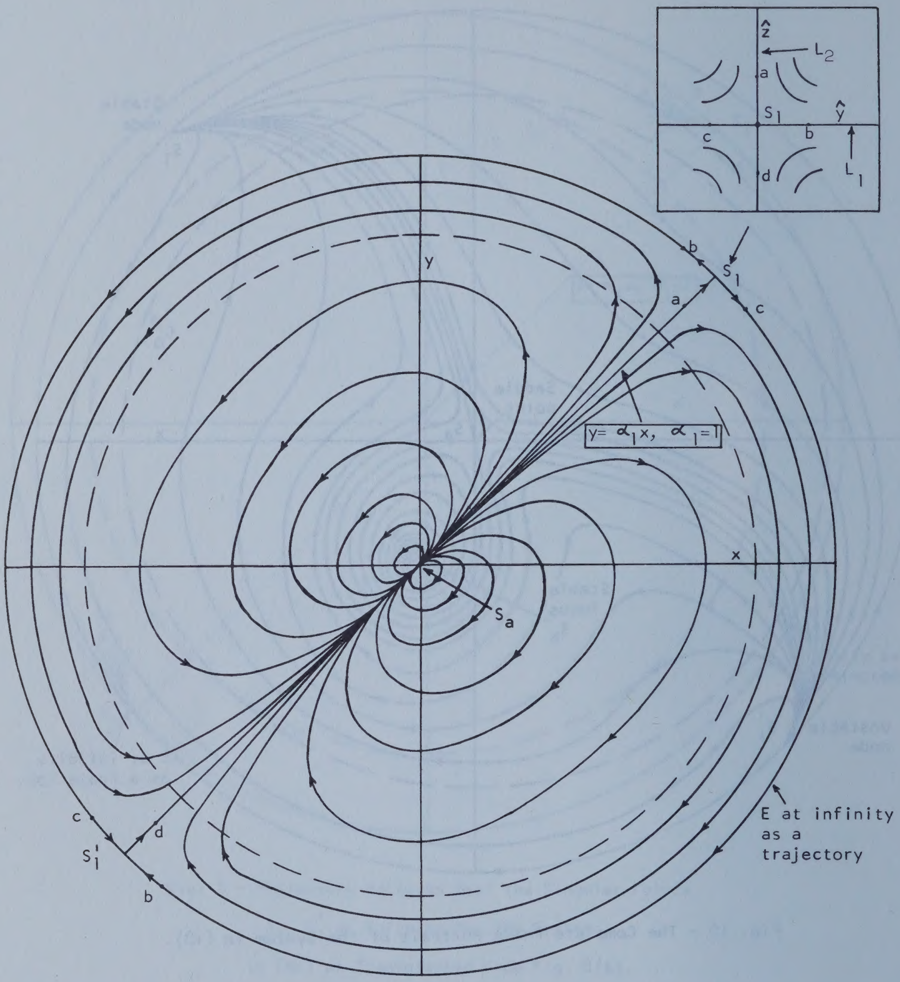

Stable
node
S_1
y
x
C_o
Saddle
point
S_a
Stable
focus
S_b
Unstable
node
S_1'
E at infinity
as a trajectory

Fig. 10 - The Complete Phase Portrait of the System in (18).



ẑ
L2
a
S1
ŷ
c
b
L1
d
b
S1
a
c
y
y= α1x, α1=1
x
Sa
c
d
S'1
b
E at infinity as a trajectory

Fig. 11 - The Complete Phase Portrait of the System in (19).

ALGEBRAIC THEORY OF LINEAR SYSTEMS*

R. E. Kalman
Stanford University
Stanford, California

1. INTRODUCTION

Elementary textbooks in system theory present the Laplace transform as a magic and mysterious gadget. Somehow the difficult problem (?) of solving linear differential equations with constant coefficients is replaced with the easy problem (?) of factoring polynomials and computing partical fractions.

In mathematics it has been long known that the solution of linear differential equations with constant coefficients is basically just a problem in linear algebra. The free differential system

$$dx/dt = Fx \tag{1.1}$$

(x = n-vector) has the general solution $x(t) = (\exp Ft)x(0)$, and of course $\exp Ft$ can be explicitly computed via the Jordan canonical form of F [1, Chap.3]. It is not necessary and not even very illuminating to solve this simple problem with the Laplace transform, which is really a rather advanced tool.** The mathematician is tempted to conclude that the cult of the Laplace transform in system theory is only explainable by fanaticism and ignorance.

Correct as this conjecture is, we must object to its proof. The natural setting for a problem in system theory is often different from the way the corresponding problem is formulated in pure mathematics. For example, instead of the free system (1.1), in system theory we consider a continuous-time, finite-dimensional, linear, constant, differential dynamical system, which in purely mathematical language is equivalent to the equations

$$dx/dt = Fx + Gu(t), \quad y(t) = Hx(t) \quad t \in T = \text{reals} \tag{1.2}$$

---

* This research was supported in part by the US Air Force Office of Scientific Research under Contract AF 49(638)-1440 with Stanford Univ.

** With respect to infinite-dimensional systems (partial differential equations, delay-differential equations, etc.) the situation is different. These systems cannot be studied by algebraic methods alone; the tools of mathematical analysis (among them, the Laplace transform) must be brought in.

where u (= input) and y (= output) are m- and p-vectors respectively. Knowing the Jordan canonical form of F is not enough to give a complete theory of (1.2). On the other hand, this system is easily analyzed by transform methods, at least when m = p = 1. So it appears that the Laplace transform is important after all, for purely mathematical reasons.

The value judgments reviewed above have been voiced for at least the past fifteen years. They are obviously subjective. As system theory is now hopefully on the way of becoming an exact science, vague personal opinions are no longer acceptable. We must not be afraid to ask the basic questions: What makes the magic work? What are the real issues irrespective of the mathematical machinery used to express them?

In this paper we shall outline a new kind of linear system theory which in our opinion gives the natural answers to these questions.

The results, already established or readily conjectured, are the following:

First, finite-dimensional linear system theory turns out to be essentially algebraic. This confirms the intuition acquired through the use of the Laplace transform as well as from the more recent "state-variable" methods which are based mainly on linear algebra.

Second, the algebraic machinery takes its most natural form in the language of modules. Classical linear algebra, which is sufficient for studying (1.1), must be generalized in an essential way to provide a natural theory for (1.2). For wholly different reasons, we confirm here a trend which has become well established in pure mathematics in the last few years [2, Book II (Algebra), Chapter 2 (Linear Algebra), Third Edition, 1962]: modules supercede vector spaces.

Third, the idea of a transform is retained only in the sense that it provides a tool for studying modules (in analogy with group characters as a means for studying finite groups).

Fourth, at long last topology enters the picture via the cohomology theory of modules. Poincaré and G. D. Birkhoff were the first to use topology to gain insight into nonlinear free differential dynamical systems. Many attempts were made after them to construct a "topological" theory of nonlinear systems, but we have learned almost nothing. Now we have a breakthrough in sight; surprisingly, it comes from algebra and linear systems and not from analysis and nonlinear systems.

Fifth, it is quite apparent that the future needs of system theory are very similar to the current research trends in homological algebra. For instance, exact sequences have natural system-theoretic interpretations, as we shall see in Sect. 6.

The ideas presented in this paper are in many ways related to the algebraic theory of finite automata [3]. That connections between linear systems and automata are natural and inevitable has been recognized already [4].

2. ABSTRACT DEFINITION OF A LINEAR DYNAMICAL SYSTEM

For ease of exposition, we consider throughout a discrete-time (finite-dimensional linear, constant) dynamical system Σ. Such a system can be described in many different ways [5]. A natural starting point is the state-variable description

$$(\Sigma^A) \qquad x(t+1) = Fx(t) + Gu(t), \quad y(t) = Hx(t), \quad t \in T = \text{integers},$$

which is analogous to (1.2), the sole difference being the time set T. The vector x is the state and $X = R^n$ is the state space.

To define a dynamical system precisely, we must specify the admissible input functions. This question is usually ignored in the literature. For us, this is important. We define the space of input functions as

$$\Omega = \{\text{all functions } u(\cdot): T \to R^m \text{ such that } u(t) = 0 \text{ for almost all } t \in T\}.$$

We can now associate an input/output description Σ^B with our system, using Σ^A, as follows: we imagine that the initial state $x(t_o)$ is zero; we apply $u(\cdot)$ and compute the corresponding output function $y_u(\cdot): T \to R^p$; we note the special property: $y_u(t) = 0$ for all t sufficiently small. So if we set

$$\Gamma = \{\text{all functions } y(\cdot): T \to R^p \text{ such that } y(t) = 0 \text{ for all sufficiently small } t\},$$

we can define

$$(\Sigma^B) \qquad f: \Omega \to \Gamma$$

by $f(u(\cdot))(t) = y_u(t)$. This function is easily computed from Σ^A, at least in principle. So Σ^A induces Σ^B.

Following a technique explored in automata theory by Krohn and Rhodes [3], we can induce an abstract description Σ^D from Σ^B. We are guided by the following kind of intuition: a dynamical system is an abstract machine in which the inputs "act" on the states. To make this notion precise, we need four essential ideas which are related to the properties of the functions f induced from Σ^A (linearity, constancy, and causality.)

(i) Linearity. If f is manufactured from the linear system Σ^A, then it is a linear function. To make this notion precise, we view Ω and Γ as R-vector spaces. Suppose t_o is the present time. By linearity of Ω, any input function $u(\cdot)$ may be written uniquely as $v(\cdot) + w(\cdot)$, were $v(t) = 0$ for all $t \geqq t_o$ and $w(t) = 0$ for all $t < t_o$. Then by linearity of f we have the decomposition

$$f(u(\cdot)) = f(v(\cdot)) + f(w(\cdot)). \tag{2.1}$$

(ii) Abstract definition of the state. We need a definition which is related directly to f.

We begin by recalling that in control theory the present state is defined semi-precisely as the least amount of information concerning the past behavior of the system which is sufficient to predict the present and future behavior provided the present and future values of the input are known. (This definition has been stressed especially by the present writer; see [6] and [7].)

In Σ^B, "past behavior" means that we know all values of $u(t)$ for all $t < t_o$. "Future behavior" means that we know $f(u(\cdot))(t)$ for all $t > t_o$. "Knowing present and future values of the input" means we know $w(\cdot)$. By the linearity relation (2.1) to know the state it suffices to know $v(\cdot)$, but some of this information may be superfluous. For if $v(\cdot)$ and $v'(\cdot)$ are two inputs such that $v(t) = v'(t) = 0$ for all $t \geqq t_o$ and if in addition

$$f(v(\cdot))(t) = f(v'(\cdot))(t) \quad \text{for all} \quad t \geqq t_o, \tag{2.2}$$

then we need not distinguish between "past behavior" due to $v(\cdot)$ or to $v'(\cdot)$. Condition (2.2) defines an equivalence relation among certain elements of Ω (write $v(\cdot) \sim v'(\cdot)$). The "least amount of information" we need about the present and past consists in knowing the equivalence class (write $[v(\cdot)]_f^{t_o}$) but not necessarily a particular element of that class.

It may seem that the above equivalence relation is rather special because it depends on t_o. In general, this is true and unavoidable, but here the situation simplifies due to constancy. (If σ^τ is the operator $\Omega \to \Omega$ defined by $u(t) \to u(t+\tau)$, then f is constant if and only if $f(\sigma^\tau u(\cdot))(t) = f(u(\cdot))(t+\tau)$ for all $\tau \in T$.) If f is constant no generality is lost in assuming that $t_o = 1$, so that the superscript may be dropped from the notation for the equivalence classes.

So we have been led to define the state space corresponding to f provisionally as the collection of equivalence classes

$$\{[v(\cdot)]_f;\ v(\cdot) \in \Omega \ \text{ and } \ v(t) = 0,\ \ t \geq 1\} = \Omega^-/\sim.$$

Since f is linear, $[0]_f$ is a subspace of Ω^-. Hence $\Omega^-/\sim\ = \Omega^-/\ [0]_f = X_f$ (or simply X) is a linear space, the state space induced by f.

(iii) Module structure. Let us examine the action of Ω on X. Being predisposed toward an algebraic point of view, we are interested in operator structures. (That is, a set A with a rich algebraic structure acts on a set B which has a more meager structure.) The correct setup turns out to be an A-module M (A = ring, not necessarily commutative; M = abelian group).* Deferring the motivation, the appropriate definitions are the following:

$M_f = \{[v(\cdot)]_f;\ v(\cdot) \in \Omega^-\}$ regarded as an abelian group, i.e., X but disregard scalar multiplication with a real number.)

$A = \{\pi(\cdot)\colon\ T \to R;\ \ \pi(t) = 0$ for $t > 0$ and also for $t < t_o$, t_o sufficiently small$\}$, with the obvious definition of addition, and with multiplication defined as convolution:

$$(\pi_1\pi_2)(t) = \sum_{\tau \in T} \pi_1(t-\tau)\,\pi_2(\tau) \tag{2.3}$$

---

[*] Recall that a module is a generalization of a vector space in that the scalars belong to a ring, not a field. Otherwise, the module axioms coincide with the vector-space axioms. See [2, loc. cit.]. In old-fashioned terminology, a module is merely an abelian group, without the operator ring.

The scalar multiplication $A \times M \to M$ is given by

$$(2.4) \qquad (\pi, [v(\cdot)]_f) \to [\pi v(\cdot)]_f$$

The reader should verify that the module axioms are satisfied by this scalar multiplication.

The definitions become a little more transparent if we observe that A is isomorphic to the ring $R[z]$ of polynomials over an indeterminate z with real coefficients, the isomorphism being given by

$$(2.5) \qquad \pi(\cdot) \sim \sum_{t \in T} \pi(t) z^{-t}$$

Thus $\pi(\cdot)$ may be regarded also as a polynomial, and (2.3) is the usual multiplication of two polynomials. Note that the shift operator σ^{τ} is now represented as z^{τ}.

(iv) *Output map and causality.* We must now check that our tentative definition of the state meets all requirements. To determine the output, we define a linear map η_f: $X \to R^p$ by

$$(2.6) \qquad y(1) = \eta_f(x(1)) = \eta_f([v(\cdot)]_f) = f(v(\cdot))(1)$$

which gives the output due to the state $x(1)$ at time $t_o = 1$. By induction, if $w(\cdot) = 0$,

$$(2.7) \qquad y(t) = \eta_f(x(t)) = \eta_f(z^{t-1}[v(\cdot)]_f) = f(z^{t-1}v(\cdot))(1) = f(v(\cdot))(t),$$

so that with our definition of the state we can indeed "predict" the present and future outputs.

This does not imply that the output of the system is a function of the state alone, because of the term $f(w(\cdot))(t)$ in (2.1). If f is *causal* (that is, $u(t) = u'(t)$ for $t < t_o$ implies $f(u(\cdot))(t_o) = f(v(\cdot))(t_o)$ then $y(1)$ is independent of $w(\cdot)$ (which is 0 for $t < 1$) and the formula $y(t) = \eta_f(x(t))$ is valid for all t. In other words, *our definition of the state is independent of the notion of causality which affects only the output map*.

We can stop now, for we have defined the desired abstract description of Σ, induced from Σ^B:

(Σ^D) $\{$an $R[z]$-module M_f and an output map $\eta_f\}$

The process of inducing Σ^D from Σ^A and Σ^B can be reversed [5]. The correspondence Σ^B and Σ^D is one-to-one [5]. However, in the sequence

$$\Sigma^A \to \Sigma^B \to \Sigma^D \to \Sigma^A \tag{2.8}$$

the first step is usually many-to-one. The process (2.8) is called *reduction*; the second Σ^A is the *reduced form* of the first Σ^A. The reduced form is unique; there is a one-to-one correspondence between Σ^D and the reduced Σ^A [5]. "Reduced" means that we are restricting our attention to the external behavior of the system, as is obvious from the first map in (2.8).

We have now all the essential ideas for the fundamental result in the algebraic theory of linear systems:

THEOREM 1. *A linear, constant, discrete-time, finite dimensional system possesses an abstract description* (M, η) *which represents the input/output behavior of the system uniquely*.

In the algebraic theory of finite automata there is a similar representation: an *automaton* is a pair $\{S, P\}$ where $S =$ semigroup, $P =$ partition.

In representing f as a pair (M_f, η_f), it is not necessary to assume that M_f is the *smallest* module which will work (called the *minimal realization of* f). But this is precisely what we have done in defining M_f as $\Omega^-/[0]_f$. We can let $M_f = \Omega^-$ since Ω^- is naturally an $R[z]$-module (with scalar multiplication $(\pi, v(\cdot)) \to \pi(\cdot)v(\cdot)$). If $A = R[z]$, Ω^- is also a free module and (Ω^-, η_f) is the *free realization of* f. (If Ω^- is not free, it may be replaced by the free module F_{Ω^-} generated by the elements of Ω^-.) The free realization contains the states $[0]_f$, which are interpreted (intuitively) as the *redundancy* caused in the free state space Ω^- by the requirement that we have a that we have a realization of f. If $M \approx \Omega^-/[0]_f$ there is no redundance and the realization is minimal.

3. SYSTEM THEORETIC INTERPRETATIONS

The mathematical concept of a module helps to clarify some of our fuzzy ideas about linear systems.

As the first example, notice that "input" in everyday usage has a double meaning: (a) a terminal at which a signal is accepted by the system; (b) the input signal itself. In module theory, an input terminal corresponds to a generator of M, while the class of acceptable input signals is the ring of operators of M. The m-input linear system (2.1) has m generators given by the m equivalence classes $\{[(1,0,\ldots,0)], \ldots, [(0, \ldots, 0, 1)]\}$. For system theory, "finitely generated" is the important case.

A module is cyclic if it is generated by a single element, that is, every state in the system has the representation $x = [\pi]$, $\pi \in A$. In system theory, we call this complete reachability; it is possible to reach any state from 0 by the application of a suitable input. Our notation shows that if $x = [\pi]$ then x is reachable by the input signal π.

If a system is capable of accepting two different classes of signals (say, AM or FM), then the module description will require two rings of operators. This is a new research topic of some interest.

M is a torsion module if for each x there is an element π_x of A such that $\pi_x x = 0$. The set of all annihilators π_x, $x \in M$ is an ideal of A, the annihilating ideal J. A very nice A is a principal ideal ring where J is generated by a single element $\chi \in A$ (write $J = (\chi)$). If the input signal χ is applied to any input terminal of such a module, the state is sent to 0. The polynomial χ turns out to be identical with the minimal polynomial of the matrix F in (2.1). "Torsion module" is equivalent to "finite-dimensional system."

It is worth noting that none of our preceding considerations depends on the underlying field of real numbers, but is valid for an arbitrary field K (for instance, integers modulo a prime), in which case $A = K[z]$, a principal-ideal domain In this case there is a complete decomposition theory which we shall outline in Sect. 5. The consideration of arbitrary fields is of practical importance because it gives us the theory of linear sequential finite-state machines, which heretofore has been developed as though it were a special branch of linear system theory.

4. EXISTENCE OF TRANSFER FUNCTIONS

Let us consider the diagram

$$\begin{tikzcd} & M_f \arrow[dr, "\nu"] & \\ \Omega^- \arrow[ur, "\mu"] \arrow[rr, "f"'] & & \Gamma^+ \end{tikzcd}$$

f is the map defined by Σ^B. μ: $v(\cdot) \to [v(\cdot)]_f$ is the canonical map which projects elements of Ω^- onto the equivalence classes defined by f. ν is the map which sends a state $[v(\cdot)]_f$ into its output sequence $y(\cdot)$, given by $y(t) = \eta_f(z^{t-1}[v(\cdot)]_f)$. The reader should verify that the diagram is commutative, that is, $f = \nu \circ \mu$. (Γ^+ is an output space such as Γ where $y(t) = 0$, $t < 1$.)

We can now claim two things [5]:

(i) the notion of "state" is intimately related to the factorization of the input/output map f as $\nu \circ \mu$;

(ii) the existence of transfer functions for finite-dimensional discrete-time linear systems is an immediate consequence of the above commutative diagram.

The first assertion is obvious.

The second assertion is in fact a new way of establishing the existence of z-transforms (for which see the works of Ya. Z. Tsypkin, J. R. Ragazzini, L. A. Zadeh, and R. H. Barker).

Observe first that Γ can be made into an $R[z]$-module in a natural way. Namely, if we write

$$y(\cdot) \approx \sum_{t \geq 1} z^{-t} y(t)$$

then Γ^+ is a subset of the formal power series ring $R^p[[z^{-1}]]$. Γ^+ becomes an $R[z]$-module (with the obvious addition) if we define the scalar multiplication as $\pi(\cdot)\cdot y(\cdot) = \hat{y}(\cdot)$, where

$$\hat{y}(t) = \sum_{0<t=\sigma-\tau} \pi(\tau)z^{\tau}y(\sigma)z^{-\sigma}$$

So the commutative diagram represents maps between $R[z]$-modules.

We can now prove the second assertion by remembering that Ω^- is finitely generated and that (by finite-dimensionality) M_f is a torsion module with annihilator χ. Let $g_1, \ldots, g_m$ be the generators of Ω^-. Their images under μ are $[g_j]_f$, $j = 1, \ldots, m$. Then $[\chi g_j]_f = 0$, and (since f is an $R[z]$-homomorphism and the diagram is commutative)

$$\chi \circ f(g_j) = f(\chi g_j) = \nu[\chi g_j] = 0.$$

So the vector power series $\chi f(g_j)$ is a polynomial. Hence $f(g_j)$ is the power series in z^{-1} obtained by formally expanding the rational fraction $(\chi f(g_i))/\chi = t_j(z)$. The $p \times m$ rational matrix $T(z)$ whose columns are the p-vectors $t_j(z)$ is the (z-transform) transfer-function matrix of our system. In other words, $t_j(z)$ (= formal power series) is the output resulting from the input sequence $v(\cdot) = 1$ applied at the j-th input terminal.

So we see that transfer functions for finite-dimensional linear systems have an essentially algebraic character.

5. DECOMPOSITION THEORY

By associating an algebraic structure with a linear system we can make use of the rich decomposition theories available in algebra. As an indication of what can be done, we consider the problem of decomposing an m-input, p-output linear system (over the real or complex field). This was fully solved by the author [8] with the aid of classical algebra, but at the expense of very complicated arguments. Using the machinery of this paper, and the sharpening of the classical results in the form given by Bourbaki [2, Book 2, Chapter 7 (1964), § 4, No. 2), this result is the following:

THEOREM 2. If M_f is a finitely generated module over a principal ideal domain (here $R[z]$), then M_f is the direct sum of finitely many cyclic modules $R[z]/(\chi_i)$, where χ_i divides χ_{i-1} and it may happen that $\chi_i = 0$ for all $i > r$. $\{(\chi_i)$ is the ideal generated by $\chi_i \in R[z]$.$\}$

If $\chi_i = 0$, then $R[z]/(\chi_i) = R[z]$ is a free cyclic module. Any free cyclic module $R[z]$ regarded as an R-vector space is infinite dimensional. This is the reason for the definition: "torsion" is equivalent to "finite-dimensional".

From Theorem 2, we obtain some interesting system-theoretic interpretations, which have been known for some time [8, 9]:

(i) The dimension of $R[z]/(\chi_i)$ regarded as an R-vector space is evidently $\delta\chi_i$ (δ = degree). So the dimension of a minimal realization of f (or of the uniquely corresponding transfer function matrix T(z)) is given by

(4.1) $$\dim X_{\min} = \Sigma\, \delta\chi_i .$$

This is a formula originally due to Brockway McMillan, who defined the left-hand side of (4.1) as the degree of the transfer-function corresponding to f.

(ii) The decomposition into cyclic modules means that the original complicated system may be thought of being composed of elementary subsystems, each completely controlled via a single input terminal. This degree of decomposition is always possible in linear systems.

(iii) The polynomials χ_i are the invariant polynomials of the square matrix F (obtained by computing the reduced Σ^A corresponding to M_f). It is possible to determine χ_i, hence the structure of F, directly from f or the transfer function matrix, without actually computing F [8].

(iv) The χ_i may be regarded also as defining the central core of a linear system; all minimal realizations can be obtained by suitably interconnecting the central core to the input and output terminals; these connections (which can be exhibited abstractly using module theory) are memoryless codes [9].

6. SYSTEM THEORY AND HOMOLOGICAL ALGEBRA

It is interesting to speculate on the interrelations between modern (homological) algebra and linear system theory. The mathematical background we must assume of the reader is elementary but of recent vintage; [11] and [12] contain readable introductory material.

Let A be a fixed ring and let M_i, $i = 0, 1, 2, \ldots$ be A-modules. (We shall concentrate on the properties of modules, without saying anything specific about A.) The sequence of modules and maps

$$(6.1) \qquad \cdots \xrightarrow{\varphi_{i+1}} M_i \xrightarrow{\varphi_i} M_{i-1} \longrightarrow \cdots \xrightarrow{\varphi_1} M_o \longrightarrow 0$$

is called exact if $\ker \varphi_{i-1} = \mathrm{im}\, \varphi_i$ for all i; it is a complex (terminology from algebraic topology) if $\varphi_i \circ \varphi_{i+1} = 0$ for all i, in other words, if $\ker \varphi_i \supset \mathrm{im}\, \varphi_{i+1}$. The homology modules of the complex are defined as $H_i = \ker \varphi_i / \mathrm{im}\, \varphi_{i+1}$. Note that "exact" is equivalent to "all homology modules are trivial."

The simpliest interesting case is the exact sequence

$$(6.2) \qquad 0 \to K \to L \to M \to 0$$

To say that it is exact means three things: (i) $L \to M$ is onto; (ii) $K \to L$ is one-to-one, and (iii) $M \approx L/K$.

The system-theoretic interpretation of an exact sequence is close at hand. Consider

$$(6.3) \qquad 0 \to [0]_f \xrightarrow{\delta} \Omega^- \xrightarrow{\epsilon} M \to 0$$

(δ = injection). If the sequence is exact then $M \approx \Omega^-/[0]_f$, in other words, M is a minimal realization of f. The homology module $H_o = M_f/\mathrm{im}\, \Omega^-$ measures the amount by which M is unnecessarily large for realizing the inputs Ω^- modulo $[0]_f$ (some states in M are not reachable via inputs in Ω^-). The homology module $H_1 = \ker \epsilon/[0]_f$ measures the degree to which M_f fails to be a realization: some elements in Ω^- (those in $\ker \epsilon$ but not in $[0]_f$) are regarded as redundant and are sent to 0 even though they may not belong to $[0]_f$.

These observations suggest a problem. Suppose we are given $[0]_f$ (the desired amount of redundancy in Ω^-) and a machine M (which is to be a minimal realization of f.) What are all possible (nonequivalent) input spaces Ω^-, in other words, what are all possible exact sequences (6.3)? The question is obviously related to the specification of the map $\Omega^- \to M$ (encoding inputs as states.)

Mathematically speaking, given the modules K and M we are to find a module L such that (6.2) is exact. Then L is called *an extension of* K *by* M. An important result of homological algebra is the description of all (nonequivalent) extensions in terms of homological concepts.

If M is a *projective* module [12], then, by definition, every exact sequence (6.3) "splits"; in other words, $\Omega^- \approx M \oplus [0]_f$. In this case Ω^- is essentially uniquely determined by M and $[0]_f$. The possibility of writing Ω^- as a direct sum of M and $[0]_f$ means that the "unobservable states" of the free realization (namely the inputs in $[0]_f$) may be "split" from the observable states (namely those of the minimal machine M). Thus "M = projective" is an abstract necessary and sufficient condition for one half of the writer's canonical decomposition theorem [13].

If M is not projective, then we consider its *free resolution*. It is the exact sequence

$$\cdots\ F_2 \xrightarrow{d_2} F_1 \xrightarrow{d_2} F_o \to M \to 0 \tag{6.4}$$

where every F_i is a free A-module. (A free resolution always exists, for we may set $F_o = F_M$ = the free module generated by M.) If the exact sequence terminates on the left as $0 \to F_n \to \ldots$, then n is the *free* (or *projective*) *dimension* of M. This dimension is unique even though the free resolution (6.4) is not. If $n_{free}(M)$ is finite, it follows by exactness that

$$M \approx F_o/F_1/ \ldots /F_n. \tag{6.5}$$

A free module is essentially the same as a vector space in that we can talk about a basis and coordinates. If M is projective, then M is a direct summand of F_M and we can use the basis in F_M to describe the elements of M. If M is not projective but $n_{free}(M) < \infty$, than by (6.5) we can still represent M as a quotient of free modules; in other words, it is possible to coordinatize M if we are willing to accept a redundant description as in (6.5).

Given an arbitrary A-module K, the free resolution of M induces the sequence

$$(6.6) \qquad 0 \to \mathrm{Hom}\ (F_o, K) \xrightarrow{d_1'} \mathrm{Hom}\ (F_1, K) \xrightarrow{d_2'} \cdots$$

where Hom (F_i, K) denotes the (A-module) of all homomorphisms $F_i \to K$, while d_i' is defined by $d_i'(\varphi) = \varphi \circ d_i$, $\varphi \in \mathrm{Hom}\ (F_{i-1}, K)$.

The sequence (6.6) is evidently a complex. Its homology modules, $H^i(M, K) = \ker d_i'/\mathrm{im}\ d_{i-1}'$ are the cohomology modules of M. (Even though the free resolution defining them is non-unique, it can be proved that the $H^i(M, K)$ are determined solely by M, K.)

It is known that the nonequivalent extensions of K by M are in one-to-one correspondence with $H^2(M, K)$. Let us give some system-theoretic flavor to this interpretation of the second cohomology module.

In view of the exact sequence $0 \to J \xrightarrow{\delta} F_M \xrightarrow{\epsilon} M \to 0$, where $J = \ker \epsilon$, any module has a representation $M \approx F_M/J$. Using the induced complex

$$0 \to \mathrm{Hom}\ (M, K) \to \mathrm{Hom}\ (F_M, K) \to \mathrm{Hom}\ (J, K) \to 0$$

it can be shown that

$$H^2(M, K) \approx \mathrm{Hom}\ (J, K)/[\mathrm{im}\ \mathrm{Hom}\ (F_M, K)].$$

Now we can give a system theoretic interpretation of H^2. Obviously $K = [0]_f$. Hom $(J, [0]_f)$ is all possible ways of encoding the available redundant states $J = \ker (F_M \to M)$ as elements of the prescribed redundant states $[0]_f$. Hom (F_M, K) is all possible ways of encoding the elements of the biggest input space $\bar{\Omega} = F_M$ as the prescribed redundant states. If Hom $(J, K) = \mathrm{im}$ Hom (F_M, K), there is no freedom; the only acceptable extension is F_M. Otherwise $H^2(M, K)$ measures the amount of leeway in the coding process and this is (intuitively) the same measuring the number of nonequivalent input spaces $\bar{\Omega}$ for which M is a minimal realization of f.

The preceding theory has a dual. Injective modules are the dual of projective modules, in strict analogy with the fact that controllability is the dual of observability [13].

REFERENCES

[1] Solomon Lefschetz, DIFFERENTIAL EQUATIONS: GEOMETRIC THEORY (book), Interscience-Wiley, 2nd Edition, 1962.

[2] N. Bourbaki, ÉLÉMENTS DE MATÉMATIQUE (book), Hermann, Paris (published in installments).

[3] K. B. Krohn and J. L. Rhodes, "Algebraic Theory of Machines", dissertation, Harvard University and MIT, 1962.

[4] M. A. Arbib, "A common framework for automata theory and control theory", J. Control (SIAM), 3 (1965), No. 2.

[5] R. E. Kalman, "Algebraic structure of finite dimensional linear systems. I. The module of Σ", Proc. Nat. Acad. Sci. (USA), Dec. 1965.

[6] R. E. Kalman, "On the general theory of control systems", Proc. 1st Internat'l Congress on Automatic Control, Butterworths, London.

[7] T. Fuller, J. Electronics and Control, circa 1962.

[8] R. E. Kalman, "Irreducible realizations and the degree of a rational matrix", SIAM J., 13 (1965) 520-544.
See also "On canonical realizations", Arch. Automatyki i Telemachaniki (Warsaw), 10 (1965) 3-10.

[9] R. E. Kalman, "On structural properties of linear constant, multivariable systems", 3rd IFAC Congress, 1966.

[10] R. E. Kalman, "On the realization of an impulse-response matrix by a differential equation", J. Mathematical Phys. (1966), to appear.

[11] D. G. Northcott, AN INTRODUCTION TO HOMOLOGICAL ALGEBRA (book), Cambridge University Press, 1960.

[12] James P. Jans, RINGS AND HOMOLOGY (book), Holt, Rinehart and Winston, 1964.

[13] R. E. Kalman, "Mathematical description of linear dynamical systems", J. Control (SIAM), 1 (1963) 152-192.

PARAMETER INVARIANCE THROUGH THE USE OF NONLINEAR COMPARATORS

JAMES M. HAM
Department of Electrical Engineering,
University of Toronto,
Canada.

ABSTRACT

Some properties of a group of feedback structures which provide input-output invariance to a single piecewise constant parameter are examined. A key feature of these structures is the use of dividers as signal comparators. The ability of some of these systems to provide invariance to constant load disturbance and to provide invariant input-output dynamics while suppressing sensor noise is illustrated. Apart from their interest as structures, these systems may prove of practical use in maintaining qualitatively useful control of some types of noisy non-minimum phase plants, and in modelling economic systems.

INTRODUCTION

Since parameters enter equations of state as multipliers of components of the state and steering vectors, it is reasonable to suppose that changes in parameters, and hence steering corrections to compensate for these changes, can be detected by performing division operations on appropriate signals in the system. In the majority of control systems signals are compared by subtraction operations. The key idea in this paper is the comparison of signals by division. It is well to remember that in economic systems ratio or per unit measures may be the primary "signals". Some possibilities of devising useful feedback control structures using ratio comparators and multiplier correctors have been considered[1,2]. This paper explores properties of the parameter invariant schemes introduced in reference 2. In some sense the systems presented are "dual" to those developed, particularly in the U.S.S.R., to achieve invariance to additive disturbance[3]. Here invariance to multiplicative parameter variation is sought.

SINGLE PARAMETER INVARIANCE

Low-Order Systems

Figs. 1, 2 and 3 show the structures of three feedback systems in which, under restricted conditions, the system response y to the input r is independent of a single parameter. In Fig.1 the input-output response is independent of gain k, in Fig.2 of pole position a and in Fig.3 of damping factor η. Each system contains a plant in which one parameter only is variable. In each system a distinctive feature is the comparison by division of a pair of signals. The divisor signal is in each case generated by a plant model. Since these systems have been briefly reported[2], the intent here is to define clearly the sense in which parameter invariance exists and to establish further properties of performance.

Consider as an example the system of Fig.1 which has as equations

$$\frac{dz}{dt} = -az + k\,m \qquad \text{(plant)} \qquad (1)$$

$$y = z + d \qquad (2)$$

$$\frac{dx}{dt} = -ax + k_d m \qquad \text{(model)} \qquad (3)$$

$$m = \left(\frac{x}{y}\right) r \qquad (4)$$

All parameters except k are constant. Let the disturbance d = 0 and k be piecewise constant with $k \neq k_d$. Suppose at t = 0, the system is in equilibrium with $r(0-) = \text{constant} \neq 0$. Then, for example, for an arbitrary step change in r at t = 0, such that $r(0+) \neq 0$, y is determined by the equation,

$$\frac{dy}{dt} = -ay + k_d r \qquad (5)$$

independently of the value of k for k = constant. Similar statements apply with respect to independence of the input-output relation in Figs. 2 and 3 when $a = \text{constant} \neq a_d$ and $\eta = \text{constant} \neq \eta_d$.

Typical evidence of the above conditions of invariance is given by the analog computer results in Fig.4. Note that when $a < a_d$ the steering signal m(t) shows "lead-network" compensation properties whereas when $a > a_d$, m(t) displays "lag-network" compensation properties. y(t) is invariant to a for $a < 0$.

High-Order Systems

In principle the foregoing conditions of invariance with respect to a single parameter are valid when this parameter is embedded in high-order linear systems including non-minimum phase systems. One of the remarkable consequences of this fact is that additive noise associated with a sensor that measures the system output y can be filtered out without influencing the input-output dynamics.

Filtering Output Sensor Noise

The technique for removing output sensor noise is suggested in Fig.5 where a low-pass filter with pole location $s = -b$ is added to the model when it is inserted after a sensor characterized by the noise $n(t)$. Fig.6(a) shows the successive reduction in fluctuation in y as the low-pass filter pole $s = -b$ approaches the origin when the plant is a non-minimum phase system having the transfer function,

$$\frac{Y(s)}{M(s)} = \frac{k\,(T_2 s-1)^2}{(T_1 s+1)\,(T_3 s+1)} \qquad \begin{array}{l} T_1 = 1 \text{ sec.} \\ T_2 = 2 \text{ sec.} \\ T_3 = 3 \text{ sec.} \end{array} \qquad (6)$$

Note that the input-output step response is invariant to the addition of the filter as the sensor noise is suppressed. Similar results can be obtained by the appropriate locating of filters in the structures of Figs. 2 and 3.

PRACTICAL PERFORMANCE

That feedback control structures employing signal dividers as distinctive elements can be devised to have some unusual properties, has been demonstrated. The author believes there are many other structures of this form. Certainly they can be extended to multivariable systems in which for example one row of a state matrix is variable. But are these systems capable of performing in a useful manner under practical conditions of parameter variation using realizable signal dividers? Circuit noise does not appear to be a problem.

Imperfect Dividers

In the system of Fig.3, both the dividend signal $z(t)$ and the divisor signal $p(t)$ go to zero when the system output is not changing. To avoid an indeterminate quotient signal $q(t)$, a cut-out threshold based on $\frac{dy}{dt}$ as suggested in Fig.7 can be used. The result of using this system in an analog simulation of Fig.3 is shown in Fig. 8. The chopping of the steering signal m(t) is caused by the rate threshold on the divider. Note that the input-output response of the system is essentially independent of η over a range of 0.1 to 2 times the model value η_d. The presence of a divisor rate threshold can of course lead to fluctuations at the threshold level.

Parameter Fluctuation

Parameter invariance is realized in the systems under discussion only when a single parameter is piecewise constant. Tests with these systems have shown that they perform in a reasonable manner when the "invariance" parameter undergoes fluctuation. Typical regulator response to a step change in pole location is shown in Fig.9, while representative regulator response to a sinusoidal variation in gain is shown in Fig.10.

Parameter Mismatch

In these systems, all plant parameters other than the "invariance" parameter are assumed to match the corresponding model parameters. In practice such a condition will not be satisfied. Fig.11 shows an effect of a plant having a pure time delay $\tau = 0.5$ seconds which is not present in the model. τ is one-half the model time-constant. The step response of y should be compared with that in Fig.4. The effect of this relatively large "unmatched" time delay leaves the settling time essentially unchanged. There is evidence[2] that momentary singularities caused by noncoincident zero-crossings of dividend and divisor signals can be clipped to obtain satisfactory results. Defining the stability boundaries of these systems in the presence of significant parameter mismatch and general changes in signal polarity poses interesting problems of analysis.

Load Disturbance

An important question is whether or not the parameter insensitive schemes being discussed are sensitive to load disturbances. It has been suggested for example that the disturbance insensitive structures associated with the "theory of invariance" are strongly parameter sensitive[4]. Fig. 12 shows a regulator response for a gain-insensitive system subjected to a sinusoidal additive disturbance d. A remarkable regulation property of the system of Fig. 1, is that the output is invariant to a constant disturbance. Equivalent behaviour in a conventional linear subtractive feedback system would require a plant with an integration.

CONCLUSIONS

The feedback structures presented show that the idea of comparing signals by ratio is feasible and generative of some interesting properties. These properties may be of practical use for example, in maintaining qualitative control of non-minimum phase plants having noisy transducers. Perhaps these structures may also prove of use in modelling economic systems in which "signals" are often fluctuations of ratios of variables. Stability boundaries for the systems described are not known and pose interesting problems of analysis in nonlinear differential equations.

REFERENCES

1. G.Lang and J. M. Ham, "Conditional Feedback", AIEE Trans., vol.74, pt.2, July 1955, pp. 152-156.

2. J. M. Ham and M. A. Hassan, "Appropriate Parameter Invariance with Nonlinear Feedback", IEEE Trans. on Automatic Control vol. AC-10, No.1, January 1965, pp. 87-89.

3. V. D. Vershinin and V. S. Kulebakin, "The Synthesis of Linear Invariant Automatic Control Systems", Automatika i Telemekhanika, vol. 26, No. 1, January 1965, pp. 41-48.

4. J. Preminger and J. Rootenberg, "Some Considerations Relating to Control Systems Employing the Invariant Principle", IEEE Trans. on Automatic Control, vol. AC-9, No.3, July 1964, pp. 209-215.

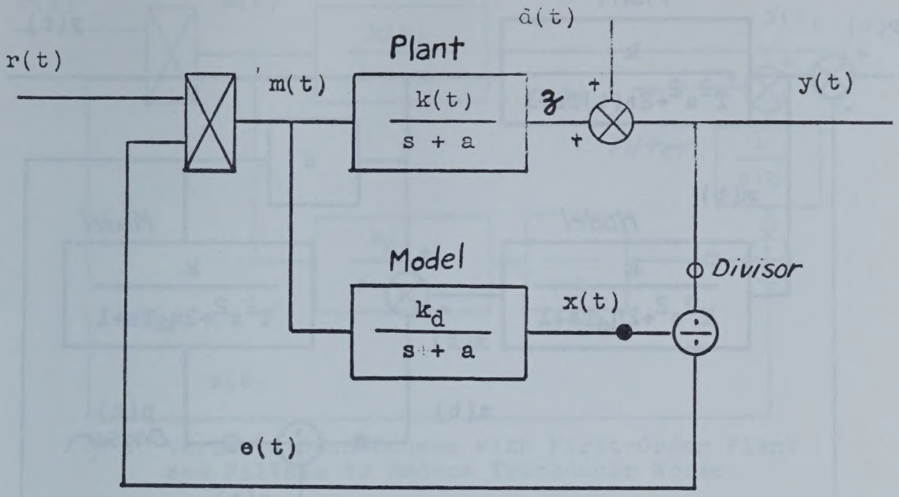

d(t)
r(t)
m(t)
Plant
k(t)
s + a
z
+
+
y(t)
Model
k_d
s + a
x(t)
Divisor
e(t)

Fig. 1 Variable-Gain Scheme with First-Order Plant

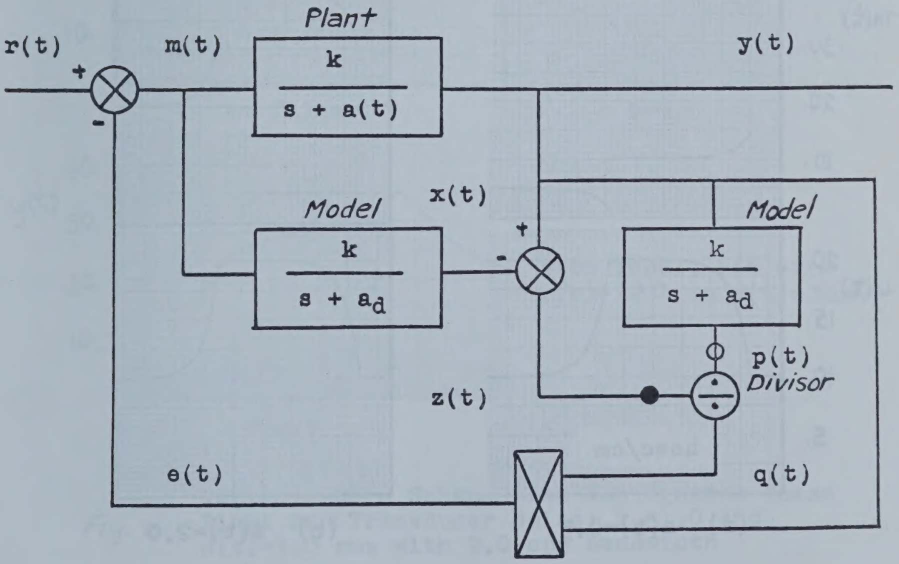

Plant
r(t)
+
-
m(t)
k
s + a(t)
y(t)
Model
k
$s + a_d$
x(t)
+
-
Model
k
$s + a_d$
p(t)
Divisor
z(t)
e(t)
q(t)

Fig. 2 First-Order Variable-Pole Scheme

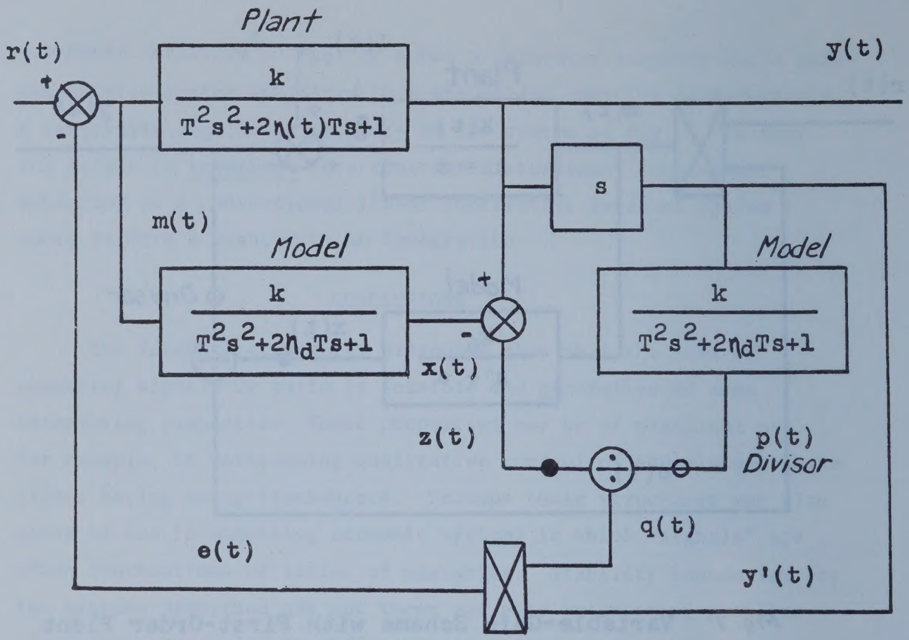
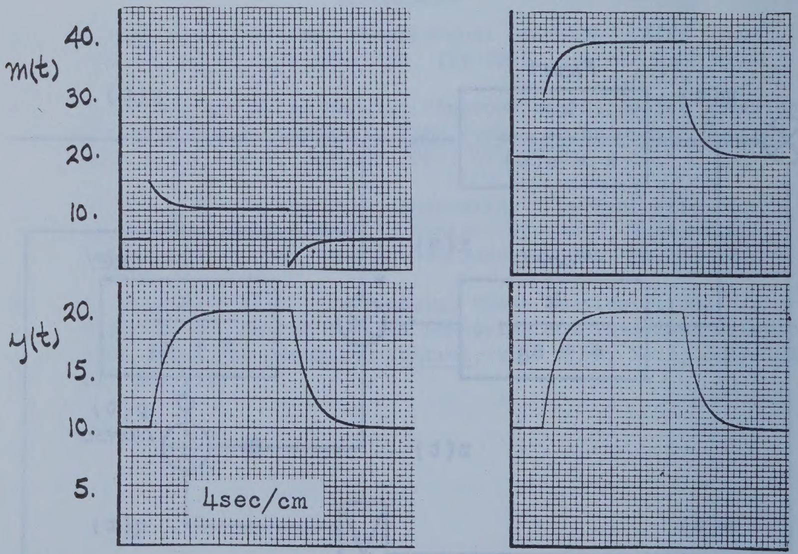

Plant
r(t)
+
-
k
$T^2s^2+2\eta(t)Ts+1$
y(t)
s
m(t)
Model
k
$T^2s^2+2\eta_dTs+1$
+
-
x(t)
Model
k
$T^2s^2+2\eta_dTs+1$
z(t)
p(t)
Divisor
q(t)
e(t)
y'(t)

Fig. 3 Variable Damping-Factor Scheme



(a) a(t)=0.5

(b) a(t)=2.0

Fig. 4 Step Response of First-Order Variable-Pole Scheme; a_d=1.0 and k=1.0

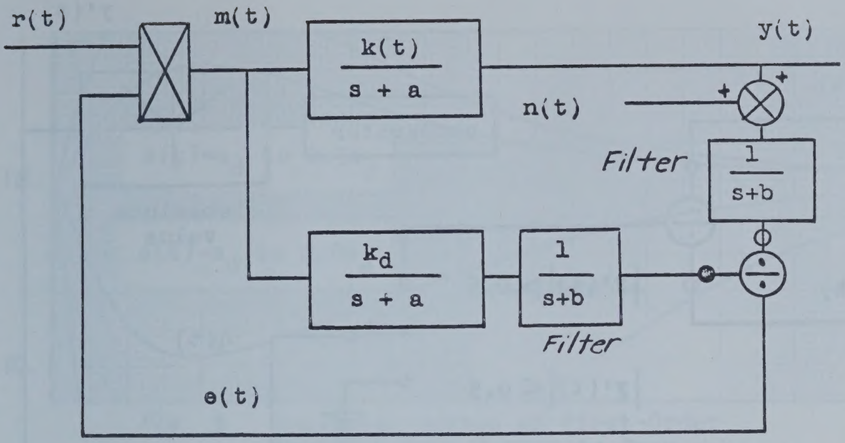
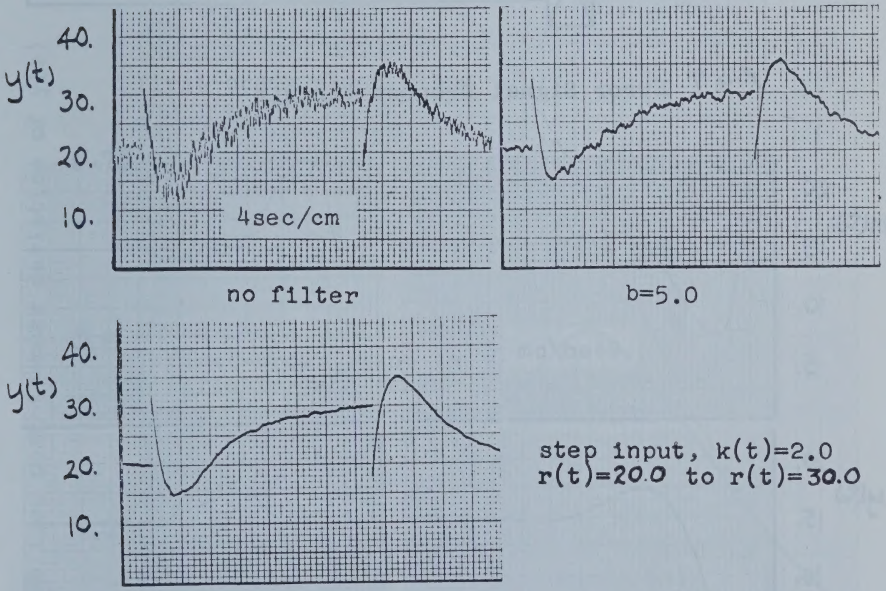

r(t)
m(t)
k(t)
s + a
y(t)
+
+
n(t)
Filter
1
s+b
k_d
s + a
1
s+b
Filter
e(t)

Fig. 5 Variable-Gain Scheme with First-Order Plant and Filters to Reduce Transducer Noise



no filter

b=5.0

b=1.0

step input, k(t)=2.0
r(t)=20.0 to r(t)=30.0

Fig. 6 Variable-Gain Scheme with Non-Minimum Phase Plant and Transducer Noise; k_d=1.0 and n(t)=1.0 rms with 2.0 cps bandwidth

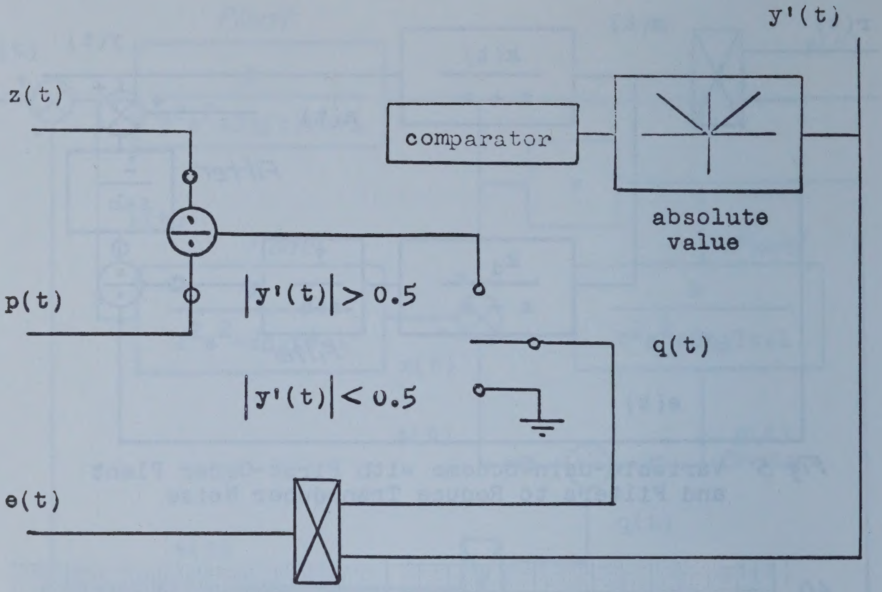
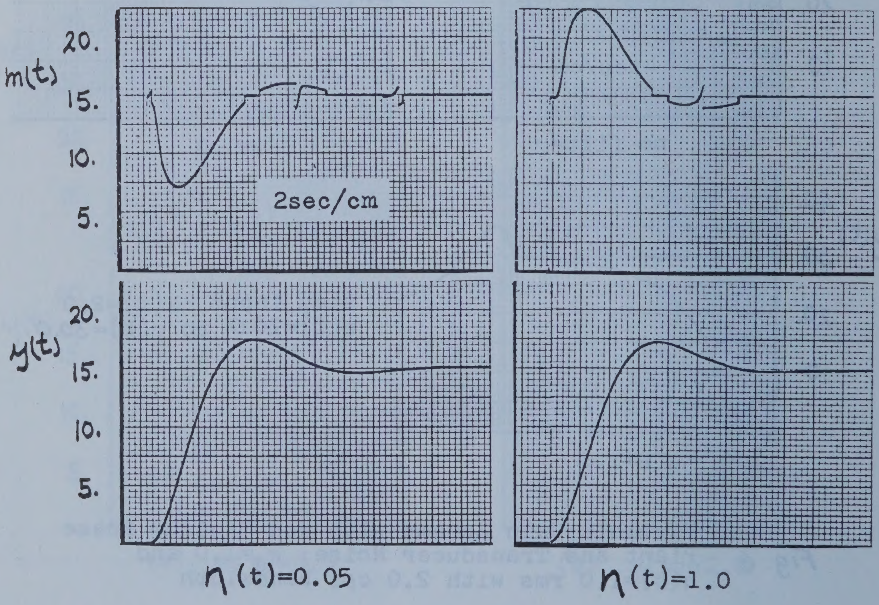

y'(t)
z(t)
comparator
absolute value
p(t)
$|y'(t)| > 0.5$
q(t)
$|y'(t)| < 0.5$
e(t)

Fig. 7 Error Signal Dropout Circuit



$\eta(t)=0.05$ $\eta(t)=1.0$

Fig. 8 Step Response of Variable Damping-Factor Scheme; k=1.0, T=1.0 and $\eta_d=0.5$

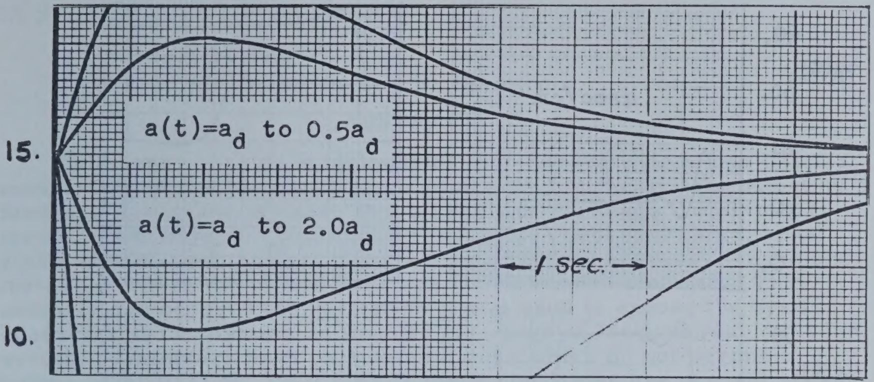


Fig. 9 Regulator Action of First-Order Variable-Pole Scheme to Step Changes in the Plant Pole; a_d=1.0, k=1.0 and r(t)=15.0

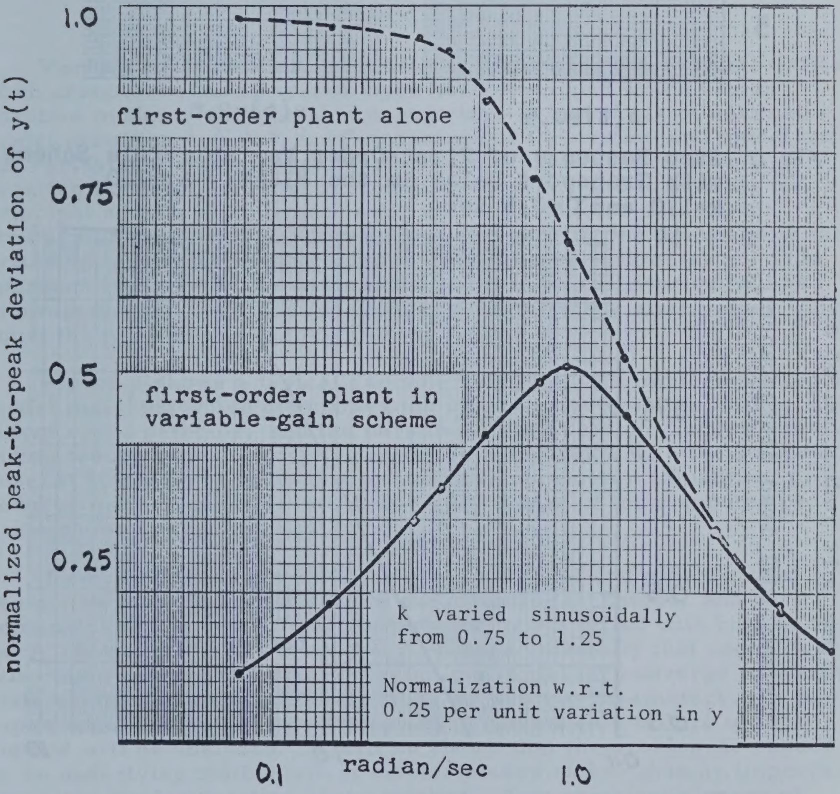


Fig 10 Frequency Response of Variable-Gain Scheme with First-Order Plant Acting as a Regulator; k_d=1.0, a=1.0 and r(t)=15.0

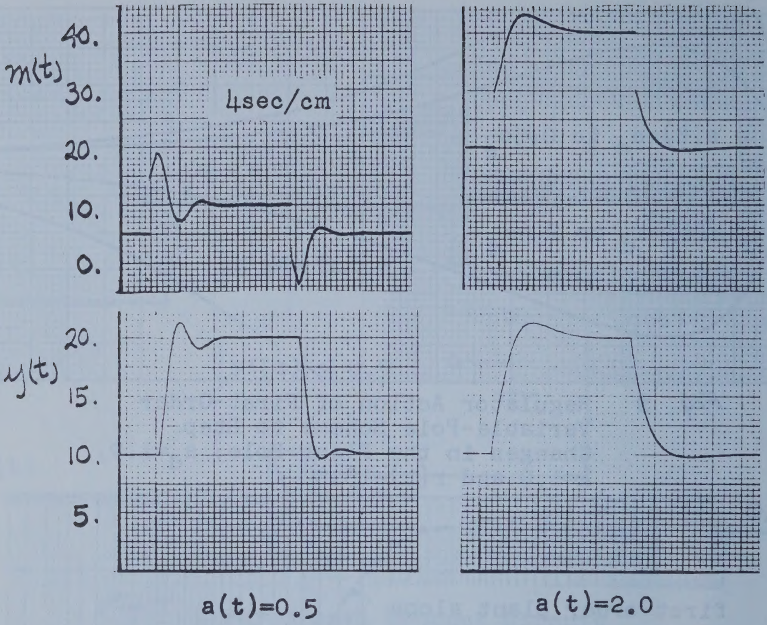


Fig. 11 Step Response of First-Order Variable-Pole Scheme with Transport Delay in the Plant; $a_d=1.0$, $k=1.0$ and $\tau=0.5$ sec

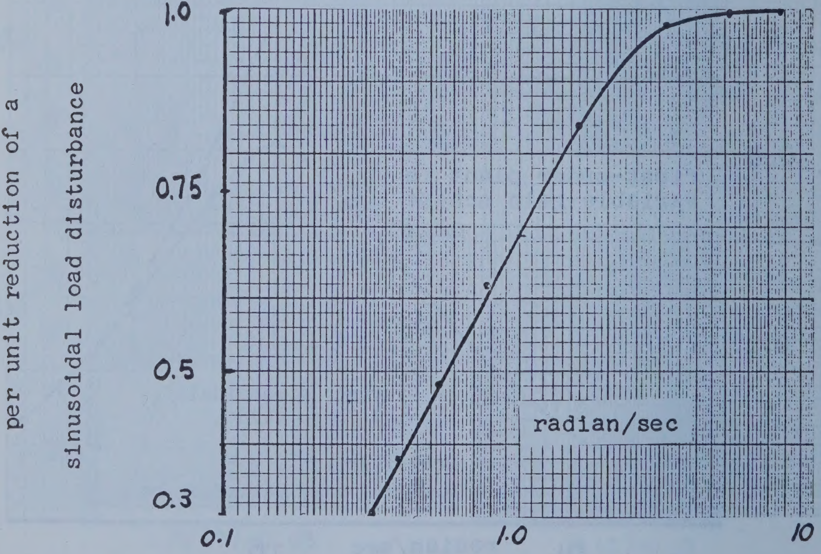


Fig. 12 Frequency Response of Variable-Gain Scheme with First-Order Plant to Load Disturbances; $k_d=1.0$, $k(t)=2.0$, $a=1.0$ and $r(t)=15.0$

ON THE PROBLEM OF INDETERMINACY IN SYSTEM IDENTIFICATION

H. F. MEISSINGER
TRW Systems, Redondo Beach, California

ABSTRACT

This paper considers methods of system identification in which a model with adjustable parameters is matched to the system output by a feedback arrangement, such that a specified non-negative error criterion is minimized. While under favorable conditions this method yields the unknown system parameters with good accuracy, there are instances where the parameters can become indeterminate. This anomaly can occur if the excitation signal used to expose the system characteristics contains an insufficient range of frequencies, or if the system parameters have particular values which do not permit a unique identification, or if both of these conditions are present at the same time. Examples of parameter indeterminacy are presented in this paper, and approaches to isolate or circumvent the underlying mathematical difficulty are explored.

INTRODUCTION

Various computational methods for identifying the characteristics of an unknown dynamic system have been developed which use an adaptive model with adjustable parameters in a process of automatic model matching [1, 2, 3, 8, 9]. Comparison of the system and model output variables yields an error term which is used to form a non-negative criterion function of modeling accuracy and to provide the principal source of information for model parameter adjustment. A strategy of steepest descent is followed in seeking the minimum of the criterion function through variation of the model parameters. This approach is based on the hypothesis that the error criterion will be minimized when the model parameters attain steady state values which equal the parameters of the unknown system.

Figure 1 shows a typical computer program for implementation of model matching which comprises the unknown system, the model, the output error detector, and the parameter optimization circuits providing feedback to the adaptive model in accordance with the steepest descent strategy. Continuous or iterative parameter adjustment procedures may be employed without major change of the computer program shown in the diagram.

There exists a substantial body of analytical and experimental results to prove the feasibility of the modeling technique, and to demonstrate that system parameters can be identified with high accuracy. However, it has been observed experimentally that under certain conditions the parameters do not converge, or converge to steady state values distinctly different from the system parameters. In this paper some anomalous conditions which can cause a failure of the method will be analyzed. It will be shown that the problem is caused by an underlying mathematical indeterminacy rather than by improper computer implementation of the method. Two principal classes of anomaly will be examined: (1) system excitation signals which are insufficient to expose the system characteristics for modeling purposes, (2) singular combinations of system parameters which do not permit the determination of a unique set of corresponding model parameters.

The possibility of encountering problems in system identification which can yield non-unique results should not be surprising. The conventional problem of analysis, viz., the problem of determining the output of a given system for a given input signal, usually has a unique solution, at least under most conditions of practical engineering interest. By contrast, the inverse problem in which the system is to be identified from known input and output signals does not necessarily have a unique solution. Cases of non-unique solutions have been demonstrated and can be analytically explained. This paper has the objective of clarifying concepts and assumptions regarding the feasibility of system identification, and provides criteria for detecting the underlying difficulty. The results have implications with respect to feasibility of plant identification in adaptive control systems.

BACKGROUND

The problem of parameter identification of a dynamic system is customarily formulated as follows. Let an n^{th} order system of differential equations in $\bar{y}$, where $\bar{y}$ is a $(n \times 1)$ vector with components y_i,

$$\dot{\bar{y}} = \bar{F}(\bar{y}, t; \bar{a}), \qquad \bar{y}(0) = \bar{y}_0 \tag{1}$$

contain m unknown parameters which are denoted by $\bar{a} = (a_1, a_2, \cdots a_m)$. The form of the system equation (1) expressed by the vector function $\bar{F}$ is assumed known. A corresponding mathematical model having the same equation form

$$\dot{\bar{z}} = \bar{F}(\bar{z}, t; \bar{\alpha}), \qquad \bar{z}(0) = \bar{z}_0 \tag{2}$$

with m adjustable parameters denoted by $\bar{\alpha} = (\alpha_1, \alpha_2, \cdots \alpha_m)$ is used in the identification procedure.* The objective is to vary the model parameters α_i simultaneously in accordance with a prescribed optimization strategy such that a specified non-negative error criterion f derived from observed differences of the state vectors $\bar{y}$ and $\bar{z}$ is minimized. Ideally, if the system and model outputs resulting from identical input perturbations $\bar{x}(t)$ are made equal by proper choice of all model parameters α_i the error criterion vanishes. In practice, the parameter adjustment is considered successful if the error criterion attains a steady state minimum near zero as the model parameters α_i reach steady state values near the correct (but unknown) parameter values a_i.

In the simplest implementation of the method the model matching error is defined by

$$e = z_1 - y_1 \overset{\Delta}{=} z - y \tag{3}$$

---

*The model is sometimes referred to as "identifier model."

which is obtained by comparing the lowest order components of the two state vectors, e.g. the system and model output quantities. Typical error criteria f(e) of the form

$$f_1(e) = \int_0^T e^2 dt \tag{4}$$

and

$$f_2(e) = e^2 \tag{5}$$

will be considered. The computer program shown in Figure 1 is designed to make simultaneous adjustments of the parameters α_i proportionately to the corresponding gradient components $\partial f/\partial\alpha_i$ of the criterion function f(e), in accordance with the method of steepest descent.

Use of the integrated error criterion (4) leads to a sequence of discrete parameter adjustment steps

$$\Delta'\alpha_i = -K\frac{\partial f_1}{\partial\alpha_i} = -2K\int_0^T e\frac{\partial e}{\partial\alpha_i}dt \tag{6}$$

applied iteratively after computing the gradient components during successive time intervals T. K is an appropriately chosen proportionality constant which determines the incremental step size. During each computation cycle the output error is measured and processed while all model parameters remain fixed.

Use of the instantaneous error criterion (5) permits continuous parameter adjustment in accordance with an adjustment law analogous to (6),

$$\dot{\alpha}_i = -K\frac{\partial f_2}{\partial\alpha_i} = -2Ke\frac{\partial e}{\partial\alpha_i} \tag{7}$$

It is advantageous in this case to introduce a lead-compensation term $q\dot{e}$ into the error criterion in order to improve the stability of the parameter adjustment process[1]. Thus

$$f_2 = (e+q\dot{e})^2 \tag{8}$$

and

$$\dot{\alpha}_i = -2K(e+q\dot{e})\left(\frac{\partial e}{\partial\alpha_i} + q\frac{\partial\dot{e}}{\partial\alpha_i}\right) \tag{9}$$

In some implementation of the continuous model matching strategy the adjustment law (7) is not mathematically rigorous since the gradient

$$\overline{\nabla f_2} = \left(\frac{\partial f_2}{\partial\alpha_1}, \frac{\partial f_2}{\partial\alpha_2}, \dots \frac{\partial f_2}{\partial\alpha_m}\right) \tag{10}$$

is defined only for error criteria derived from models with fixed parameters[3]. However, on approaching the vicinity of the correct parameter values a_i the adjustment rate $\dot{\alpha}_i$ becomes sufficiently small so that the descent to the minimum of the criterion function f_2 approximately follows a path that corresponds to the gradient $\overline{\nabla f_1}$ for fixed model parameters.

Computer implementation of the adjustment laws (6) or (7) is facilitated by the use of sensitivity coefficients of the error term e , which are defined as partial derivatives with respect to the parameters α_i,

$$\frac{\partial e}{\partial \alpha_i} = \frac{\partial z}{\partial \alpha_i} \triangleq u_i \tag{11}$$

These terms are obtainable by simultaneous solution of a set of auxiliary differential equations known as sensitivity equations[4], or by finite difference approximation, $\Delta z / \Delta \alpha_i$. For linear systems with constant coefficients one auxiliary equation often suffices to generate the m function $u_1, u_2, \ldots u_m$ required for determination of m unknown parameters[5, 6].

Properties of iterative and continuous parameter optimization processes using steepest descent strategies of the type defined by equations (6) and (7) have been extensively investigated[3, 7, 8, 9]. Stability and convergence criteria have been derived for special cases, notably for iterative optimization[8].

Results obtained from experimental parameter identification studies lead to the following observations which are supported by theory[3]. The error criterion can be effectively minimized by a well-designed parameter optimization circuit such that accurate matching of system and model output variables is achieved. However, residual steady state errors of individual parameters cannot always be detected and hence may remain uncorrected. Thus, a distinction must be made between "external accuracy" expressed in terms of residual model matching error and "internal accuracy" expressed in terms of residual parameter error. Attainment of equally high internal and external accuracy cannot be assured. The following error sources contribute to the residual parameter error:

(1) Presence of parasitic noise in the system output, not matched by the identifier model.

(2) Simplified model equation which neglects higher order terms or nonlinear elements in the system.

(3) Inaccurate computer implementation of the model equation, inaccurate detection of the model matching error, inaccurate parameter adjustment circuits (bias and scale factor errors), etc.

(4) Shallow minimum of the criterion function.

Parasitic noise and unmatched terms in the system equation generally cause a fluctuation of steady state parameter values. Bias terms in the model parameter implementation tend to cause fixed offsets of the parameters obtained. (Inaccuracy in the sensitivity coefficients affects the adjustment path rather than the steady state parameter values and is only of secondary importance.) Parameter accuracy can be improved by refinement of the model structure; by skillful computer programming and scaling; and by redesign of the parameter adjustment circuits, such that the sensitivity of the error criterion near the minimum is increased. A selective increase of the adjustment gain in parameter circuits that operate on inherently small sensitivity coefficients will also often prove helpful.

INSUFFICIENCY OF EXCITATION SIGNAL

Frequency-Dependence of Parameter Determination Accuracy

The subsequent discussion is concerned with parameter accuracy and determinacy problems which are inherent in system identification, rather than with merely computer-induced causes of inaccuracy.

If the excitation signal x(t) used as input to the system and model equations is insufficient to produce significant responses over a frequency range encompassing the essential system characteristics it may be impossible to identify some or all system parameters by the model matching method. Extensive experimental studies have shown that random-appearing or genuinely random input signals with frequency ranges encompassing the bandwidth of the system in question will, in general, yield well-defined parameters of good accuracy. The accuracy can be ascertained by testing the parameter-optimization circuits against a system with known parameters. This is illustrated by experimental results from a study involving a known second-order linear system with four parameters given by the equation

$$\ddot{y} + a_1 \dot{y} + a_2 y = a_3 \dot{x} + a_4 x \tag{12}$$

or by the equivalent transfer function

$$G_s(s) = \frac{Y}{X} = K_0 \frac{\tau_1 s + 1}{s^2 + 2\xi \omega_2 s + \omega_2^2} \tag{13}$$

Parameter identification was performed by using as excitation signal the output of a gaussian noise generator which was first passed through a third order low-pass filter with various cutoff frequencies. Three systems with frequency characteristics as illustrated in Figure 2 were used in this parameter determination experiment. The systems differ in terms of the location of their real-valued poles and zeros, with systems 2 and 3 having a dominant lead time constant, and system 3 having a double pole. Errors observed in the resulting parameter values K_0, τ_1, ω_2 and ξ (see Table) indicate that the accuracy of parameter determination deteriorates sharply as the cutoff frequency ω_c is reduced from 20 to 0.4 rad/sec, i.e., significantly below the system bandwidth. Systems 2 and 3 exhibit the effect most clearly.

In order to explain the frequency-dependence of parameter determination accuracy it is useful to examine the relative sensitivities of individual parameters α_j and α_k, expressed by the ratio of the respective sensitivity coefficients

$$r_{jk} = \frac{u_j}{u_k} \tag{14}$$

Consider, for example, a first order, three-parameter system (or model) under sinusoidal excitation having the transfer function

$$G(s) = \frac{Z(s)}{X(s)} = K_0 \frac{\tau_1 s + 1}{\tau_2 s + 1} \tag{15}$$

where $K_0 = 2$, $\tau_1 = 0.5$ sec, $\tau_2 = 0.1$ sec.

The sensitivity of z(t) with respect to the parameters K_o, τ_1, τ_2 can be expressed by

$$u_o \triangleq \frac{\partial z}{\partial K_o}, \qquad u_1 \triangleq \frac{\partial z}{\partial \tau_1}, \qquad u_2 \triangleq \frac{\partial z}{\partial \tau_2} \tag{16}$$

or by the transforms U_o, U_1, U_2 obtained from (15)

$$\frac{U_o}{X} = \frac{\tau_1 s + 1}{\tau_2 s + 1}, \qquad \frac{U_1}{X} = \frac{K_o s}{\tau_2 s + 1}, \qquad \frac{U_2}{X} = -\frac{K_o s(\tau_1 s + 1)}{(\tau_2 s + 1)^2} \tag{17}$$

Parameter Determination Accuracy as a Function of Excitation Signal Bandwidth

System	Bandwidth ω_c (rad/sec)	Parameter Error (Percent)			
		K_o	τ_1	ω_2	ξ
1	20	-	-	-	-
	3	-31.8	+6.0	+132	-26.8
	0.4	-9.8	-46.4	+136	-25.1
2	20	+1.3	-1.5	+0.9	+3.0
	3	-0.3	-1.0	-0.7	-5.2
	0.4	+0.7	-11.5	-11.6	-40.3
3	20	+0.7	+1.5	-0.8	-3.0
	3	0	0	+0.3	-5.6
	0.4	-1.0	-0.5	+1.3	-10.7

Bode diagrams of the amplitudes of U_o/X, U_1/X and U_2/X are plotted in Figure 3. The sensitivity ratios expressed in transfer function form thus become

$$R_{10} = \frac{U_1}{U_0} = \frac{K_o s}{\tau_1 s + 1}, \qquad R_{20} = \frac{U_2}{U_0} = -\frac{K_o s}{\tau_2 s + 1}, \qquad R_{21} = \frac{U_2}{U_1} = -\frac{\tau_1 s + 1}{\tau_2 s + 1} \tag{18}$$

Their Bode diagrams are also shown in Figure 3. It is evident that in the lower frequency region the amplitudes of u_1 and u_2 are much smaller than those of u_o. The relative magnitude of sensitivities determines the relative accuracy with which the corresponding parameters can be identified[3,6]. Consequently, from the Bode diagrams of Figure 3 it can be concluded that with very low excitation frequencies only K_o can be determined with good accuracy. As the excitation frequency is increased to values $1/\tau_1$ and $1/\tau_2$ the amplitudes of u_1 and u_2 increase and the parameter determination accuracy of τ_1 and τ_2 becomes comparable to K_o. For the example shown the parameters τ_1 and τ_2 become more sharply defined than K_o if excitation frequencies increase much above the bandwidth of the system. These results derived for relative parameter sensitivities in the case of sinusoidal excitation can be readily extended to the case of random excitation, by considering the effect of various cutoff frequencies ω_c on the rms-values of the parameter sensitivities. The frequency response shown in Figure 3 indicates which sensitivities dominate in the rms sense at various excitation bandwidths. A similar analysis carried out for the second-order equation (13) explains the relative parameter accuracy obtained experimentally in the above example.

Parameter Determinacy Problems Associated With Sinusoidal Excitation

Consider again the example of the four-parameter system (12) discussed in the preceding section. Comparison of the system equation and the model equation

$$\ddot{z} + \alpha_1 \dot{z} + \alpha_2 z = \alpha_3 \dot{x} + \alpha_4 x \tag{19}$$

yields the relation

$$(\ddot{y} - \ddot{z}) + (a_1 \dot{y} - \alpha_1 \dot{z}) + (a_2 y - \alpha_2 z) - (a_3 - \alpha_3)\dot{x} - (a_4 - \alpha_4)x = 0 \tag{20}$$

For the case of ideal model matching, where e = z - y ≡ 0, equation (20) reduces to an algebraic relation between the parameter errors $\Delta\alpha_i = a_i - \alpha_i$, viz.

$$\Delta\alpha_1 \dot{y} + \Delta\alpha_2 y - \Delta\alpha_3 \dot{x} - \Delta\alpha_4 x = 0 \tag{21}$$

which must be satisfied identically, i.e., at all times t by the steady state parameter values α_i. For correct parameter values all $\Delta\alpha_i$ are zero. Hence (21) is satisfied regardless of the character of the excitation signal. It is seen that at least for random excitation x(t) no other parameter combination can satisfy (21) identically. Therefore, only the correct parameters can be obtained by the identification process in this case.

If a pure sinusoidal excitation $x(t) = A \sin \omega t$ is used, the steady state output of the system is also sinusoidal, $y(t) = B \sin(\omega t + \varphi)$, and equation (21) may be separated into

$$\begin{aligned} B\omega \cos\varphi\, \Delta\alpha_1 + B \sin\varphi\, \Delta\alpha_2 - A\omega\, \Delta\alpha_3 &= 0 \\ B\omega \sin\varphi\, \Delta\alpha_1 - B \cos\varphi\, \Delta\alpha_2 + A\, \Delta\alpha_4 &= 0 \end{aligned} \tag{22}$$

by collecting coefficients of sin ωt and cos ωt respectively. This underdetermined algebraic system admits infinitely many (false) solutions $\Delta\alpha_i \neq 0$ in addition to the correct solution $\Delta\alpha_i = 0$, i.e., the parameters become indeterminate. It can be concluded that not more than two parameters will be determined uniquely when the system is excited by a single sinusoid. This result agrees with observations made in experimental parameter identification studies. Additional discussion of this type of indeterminacy may be found in a study by Bekey[10]. Miller[11] has extended these results by showing that the excitation signal must contain at least n nonharmonic sinusoidal components to permit unique identification of 2n parameters.

Parameter indeterminacy of a different type can occur with system excitation by a pure sinusoid if the form of the identifier model is inappropriately chosen, i.e., not of the same order as the system in question. The following example serves to illustrate this case. Let the system have a second-order transfer function of the form (13), with $\tau_1 = 0$. Parameter identification is attempted by means of a model of first order,

$$G_m(s) = \frac{Z}{X} = \frac{K_1}{s+\omega_1} \tag{23}$$

A sinusoidal input signal x(t) of frequency ω produces identical steady state output signals y and z if

$$K_0 \omega_1 = K_1 (\omega_2^2 - \omega^2) \tag{24}$$

$$K_0 = 2K_1 \zeta \omega_2 \tag{25}$$

This follows from equating $G_m(s)$ and $G_s(s)$. Hence, with an arbitrarily chosen excitation frequency ω one can achieve model matching by adjusting the transfer function parameters K_1, ω_1 in accordance with (24) and (25). Conversely, model matching can be achieved by adjusting only the parameter K_1 in accordance with (25) but assigning an arbitrary value to ω_1 provided the excitation frequency ω is selected so as to satisfy equation (24). Ambiguity of parameters in this example results from the structural mismatch of system and model equations as well as from the insufficiency of system excitation by a pure sinusoid.

Determinacy problems illustrated by the foregoing examples can be resolved by introducing an excitation signal of adequate frequency content, e.g., a composite periodic waveform or a random noise signal. The indeterminacy discussed in connection with equation (21) is readily eliminated in this manner. An incorrect model structure, as in (23), will be exposed through nonharmonic or random excitation since this will cause the parameters to fluctuate continuously rather than to attain constant steady state values. The error criterion is not necessarily a good indicator of incorrect model structure because a high-gain parameter-tracking circuit is capable of producing consistently small terminal output errors after the initial parameter adjustment has taken place.

The example illustrated in Figure 4(a) shows an attempt of adapting a first-order model

$$\dot{z} + \beta_1 z = \beta_2 \dot{x} + \beta_3 x \tag{26}$$

to a higher-order system. A random excitation signal was used. The unknown "system" in this case was a human operator performing a manual control task. In such a task it is known that the operator's input-output response is at least of second order. The model parameters β_i in Figure 4(a) show substantial fluctuations while the model matching error e is negligible. Figure 4(b) shows the result obtained with a reduced parameter adjustment gain. The results now exhibit a more noticeable mismatch while the parameters show less fluctuation.

Actually, it is known from comprehensive studies of manual control responses that the lead compensation parameter β_2 should be substantially positive rather than zero, as shown in Figure 4(b). The time lag parameter $1/\beta_3$ is known to be at least two to three times larger than the value 0.03 sec obtained by the parameter tracker. This example illustrates the importance of judicious choice of the model structure and of cautious interpretation of the parameter values obtained.

SINGULAR PARAMETER COMBINATIONS

Parameter indeterminacy problems of a different type can result if the system parameters to be identified assume certain unfavorable combinations which will be termed "singular." The second-order, four-parameter system (12) with random noise excitation will again be considered to illustrate this anomaly. The system transfer function

$$G_s(s) = \frac{Y}{X} = \frac{a_3 s + a_4}{s^2 + a_1 s + a_2} \tag{27}$$

may be approximated by

$$G_s'(s) = \frac{a_3 s + a_4}{a_1 s + a_2} \tag{28}$$

if the noise excitation signal has a low cutoff frequency ω_c compared to the system bandwidth. If the system parameters are related by

$$\frac{a_3}{a_1} = \frac{a_4}{a_2} = C_1 \tag{29}$$

(where C_1 is the zero-frequency gain of G_s) the system behaves approximately as if G_s were frequency-independent. The model parameters α_1, α_2, α_3, α_4, must assume corresponding ratios α_3/α_1, α_4/α_2. The computer determines parameters α_i which satisfy this condition but do not necessarily match the individual system parameters a_i, e.g., they can fluctuate or exhibit monotonic drift.

Condition (29) signifies a singular combination of parameter values. Such a condition is illustrated by an iterative sequence of four parameters obtained experimentally in a second-order model of the form (27), as shown in Figure 5. A plot of the parameter ratios α_3/α_1 versus α_4/α_2 (Figure 6) indicates that the sequence moves rapidly from an arbitrarily chosen initial point P_0 to a stationary end condition P_5 near the locus of singular parameter combinations. In accordance with (29) this locus is the line of unit slope through the origin. Like the previous example (Figure 4) this case was taken from a study of mathematical models of human control operators. The random excitation signal had a cutoff frequency of 0.3 rad/sec while the "bandwidth" of the human operator was of the order of 3 rad/sec. Hence (28) is a valid approximation and explains the lack of definition of the four individual parameters. It should be noted that this result represents an exceptional case which has been selected for illustration purposes. More typical four-parameter models of the human operator obtained for a variety of manual control tasks do not exhibit this anomaly. The indeterminacy, caused by unfavorable parameter values in combination with a low excitation signal bandwidth, could have been circumvented by using a higher frequency of excitation.

An indeterminate condition caused by singular parameter values, independent of the excitation bandwidth, is illustrated by the three-parameter system

$$G(s) = \frac{b_2 s + b_3}{s + b_1} \tag{30}$$

which degenerates to $G = C_o$ if the parameters assume values such that $b_2 = b_3/b_1 = C_o$. In this case b_1 and b_3 cannot be identified individually since the system response now depends simply on the ratio b_3/b_1. Sensitivity coefficients $\partial z/\partial \beta_1$, $\partial z/\partial \beta_3$ can be obtained formally by solving the respective sensitivity equations, as in the nonsingular case, but they become meaningless in the vicinity of the singular condition $\beta_3/\beta_1 = C_o$. In practice, a physically realizable system response cannot be described by G = constant for arbitrarily large excitation frequencies, but would be attenuated by a time-lag term not exhibited in (30).

DISCUSSION AND CONCLUSION

The examples of parameter indeterminacy investigated in this paper have been traced to conditions of inadequate exposure of system characteristics by excitation signals of limited frequency range, to singular sets of system parameters, or to a combination of these causes. Granted that these anomalous conditions are not frequently encountered, it is necessary for the investigator to guard against their occurrence and to recognize their symptoms. Unawareness of the underlying mathematical causes may lead to a fruitless search for computer programming errors or equipment malfunctions when, in fact, the circuits are functioning satisfactorily.

Considerable circumspection is required on the part of the investigator to detect conditions where the parameters converge to incorrect steady state values for which the error criterion attains a seemingly satisfactory minimum. Such cases are possible if the chosen model structure is grossly inadequate. It is also important to detect conditions of poor parameter accuracy which will occur if the system operates in the vicinity of an indeterminacy region of the parameter space.

In summary, the following steps will be useful to detect and remedy underlying conditions which cause indeterminate or semi-determinate model parameters:

(a) Determine if the frequency range of the excitation signal is sufficient for accurate parameter identification. When in doubt increase the excitation bandwidth and observe if this affects the model parameters obtained in steady state.

(b) If normal system operation precludes wideband random excitation, accept the resulting parameters with caution, particularly those parameters characterizing high-frequency response of the system.

(c) Introduction of additional dither signals can serve to expose system responses more adequately in order to improve parameter determination accuracy.

(d) In case of systematic drift of several or all parameters check the assumed model equation for possible singular parameter values. If a singular condition is present try to modify the model structure to circumvent the singularity, e.g., choose a model with fewer parameters, introduce parameter ratios, etc. In addition, attempt to improve parameter determinacy by increasing the bandwidth of the excitation signal.

Determinacy problems of the type discussed in this paper have been encountered by other investigators working in the fields of system identification, parameter optimization, and adaptive control. Detection and correction of underlying mathematical conditions causing indeterminacy has proved important in practical system identification studies conducted by the author. It is apparent that mathematically induced determinacy problems will similarly affect other approaches to system identification than the model adjustment approach presented here. Further study is required to develop a systematic theory of modeling feasibility and to establish criteria for uniqueness and existence of solutions.

ACKNOWLEDGMENT

The author wishes to thank G. A. Bekey, Associate Professor of Electrical Engineering of the University of Southern California, E. P. Todosiev and R. E. Rose, Members of the Technical Staff of TRW Systems, for their participation in conducting the studies to which this paper is related. This work was sponsored by NASA Contract, NAS 1-4419, awarded by Langley Research Center, under a study program on human operator models.

REFERENCES

1. Margolis, M., "On the Theory of Process Adaptive Control Systems, the Learning Model Approach," Ph.D. Dissertation, University of California at Los Angeles, June 1960, and Report AFOSR TN 60-618, May 1960.

2. Potts, T. F., Ornstein, G. N., and Clymer, A. B., "The Automatic Determination of Human and Other System Parameters," Proc. Western Joint Computer Conference, Los Angeles, May 1961, pp. 645-660.

3. Meissinger, H. F., and Bekey, G. A., "An Analysis of Continuous Parameter Identification Methods," scheduled for publication in SIMULATION, February 1966.

4. Meissinger, H. F., "The Use of Parameter Influence Coefficients in Computer Analysis of Dynamic Systems," Proc. Western Joint Computer Conference, pp. 181-92, May 1960.

5. Tomović, R., "Sensitivity Analysis of Dynamic Systems," McGraw-Hill, New York, 1964.

6. Bekey, G. A., Meissinger, H. F., and Rose, R. E., "A Study of Model Matching Techniques for the Determination of Parameters in Human Pilot Models," NASA Contractor Report CR-143, January 1965.

7. Bekey, G. A., and McGhee, R. B., "Computing Methods in Optimization Problems," edited by A. V. Balikrishnan and L. W. Neustadt, Academic Press, 1964, pp. 305-327.

8. McGhee, R. B., "Identification of Nonlinear Dynamic Systems by Regression Analysis Methods," Ph.D. Dissertation, University of Southern California, June 1963.

9. Eykhoff, P., "Some Fundamental Aspects of Process-Parameter Estimation," IEEE TRANSACTIONS, October 1963, Vol. AC-8, No. 4, pp. 347-357.

10. Bekey, G. A., "Continuous Identification of the Parameters of Space Vehicle Dynamics," Interim Technical Report No. SSD-TR-65-116, USCEE Report 138, August 1965.

11. Miller, B. J., "A Theoretical and Analog Study of a Steep Descent Coefficient Computer for Process Analysis and Adaptive Control," Ph.D. Dissertation, Ohio State University, 1962.

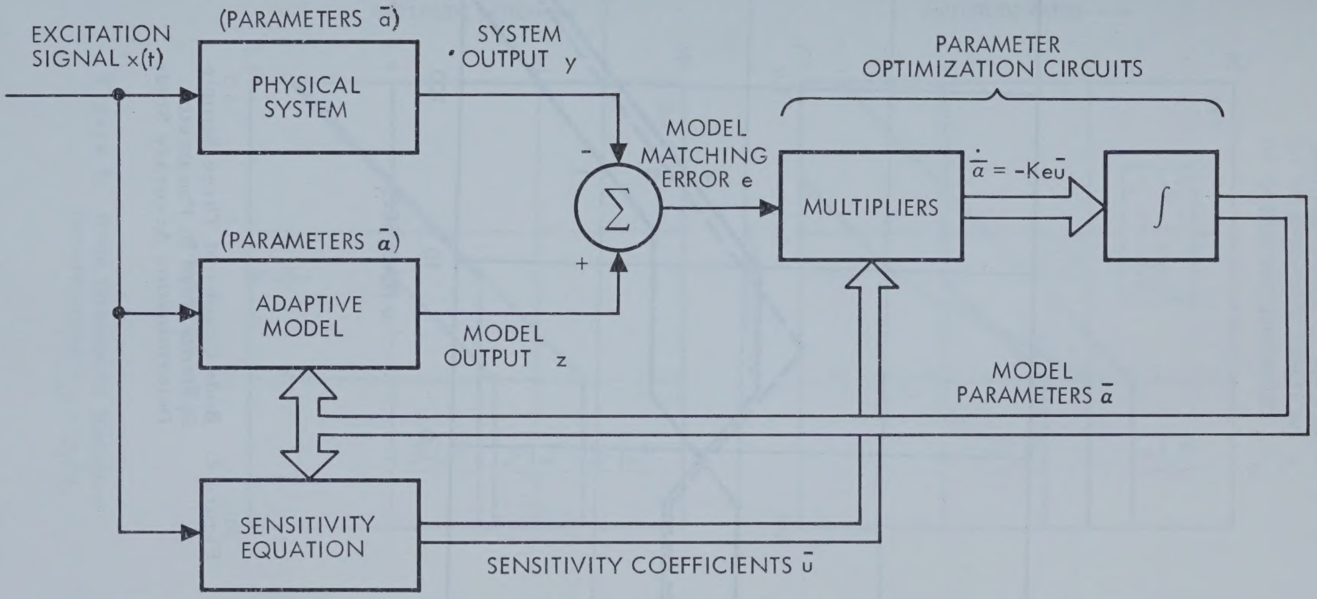

EXCITATION SIGNAL x(t)
(PARAMETERS $\bar{a}$)
PHYSICAL SYSTEM
SYSTEM OUTPUT y
(PARAMETERS $\bar{\alpha}$)
ADAPTIVE MODEL
MODEL OUTPUT z
SENSITIVITY EQUATION
-
+
Σ
MODEL MATCHING ERROR e
PARAMETER OPTIMIZATION CIRCUITS
MULTIPLIERS
$\dot{\bar{\alpha}} = -Ke\bar{u}$
$\int$
MODEL PARAMETERS $\bar{\alpha}$
SENSITIVITY COEFFICIENTS $\bar{u}$

Figure 1. Block Diagram of System Identification by Adaptive Identifier Model

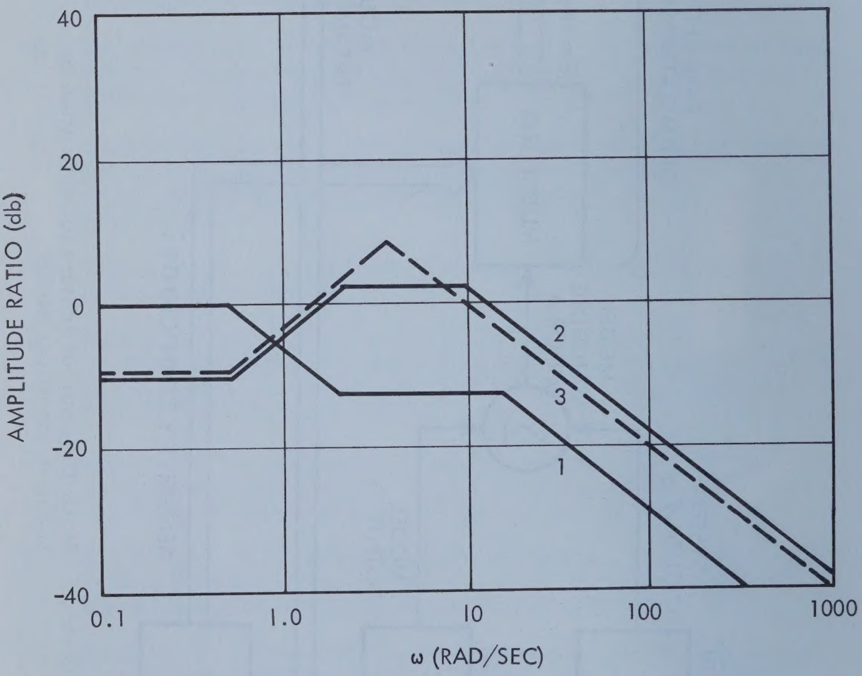


Figure 2. Bode Graphs of Three Sample Systems Used in Parameter Determination Accuracy Study

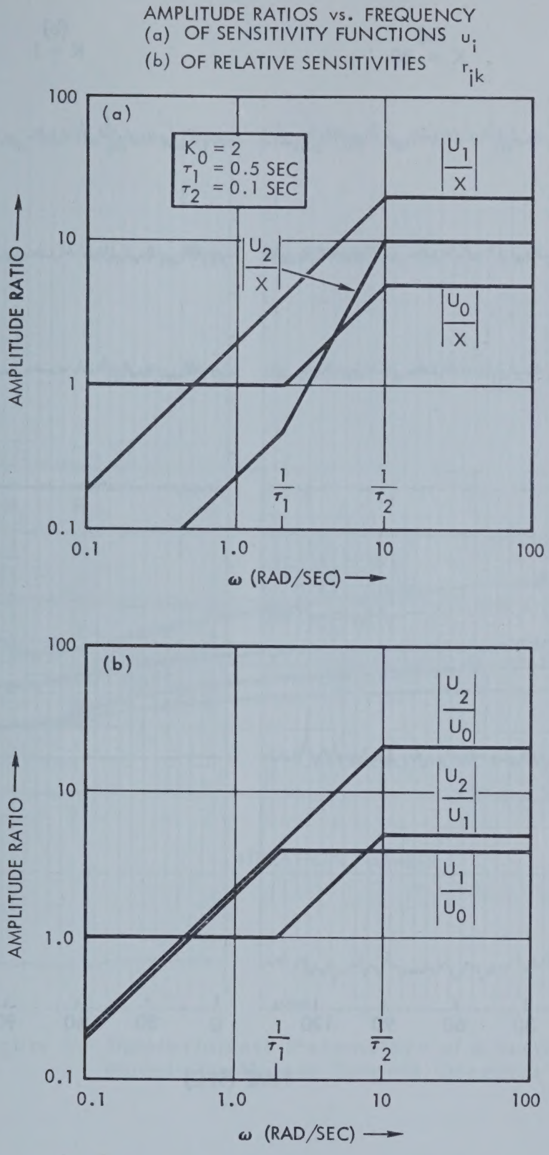

AMPLITUDE RATIOS vs. FREQUENCY
(a) OF SENSITIVITY FUNCTIONS u_i
(b) OF RELATIVE SENSITIVITIES r_{jk}

Figure 3. Bode Graphs of Relative Sensitivities $r_{jk} = u_j/u_k$

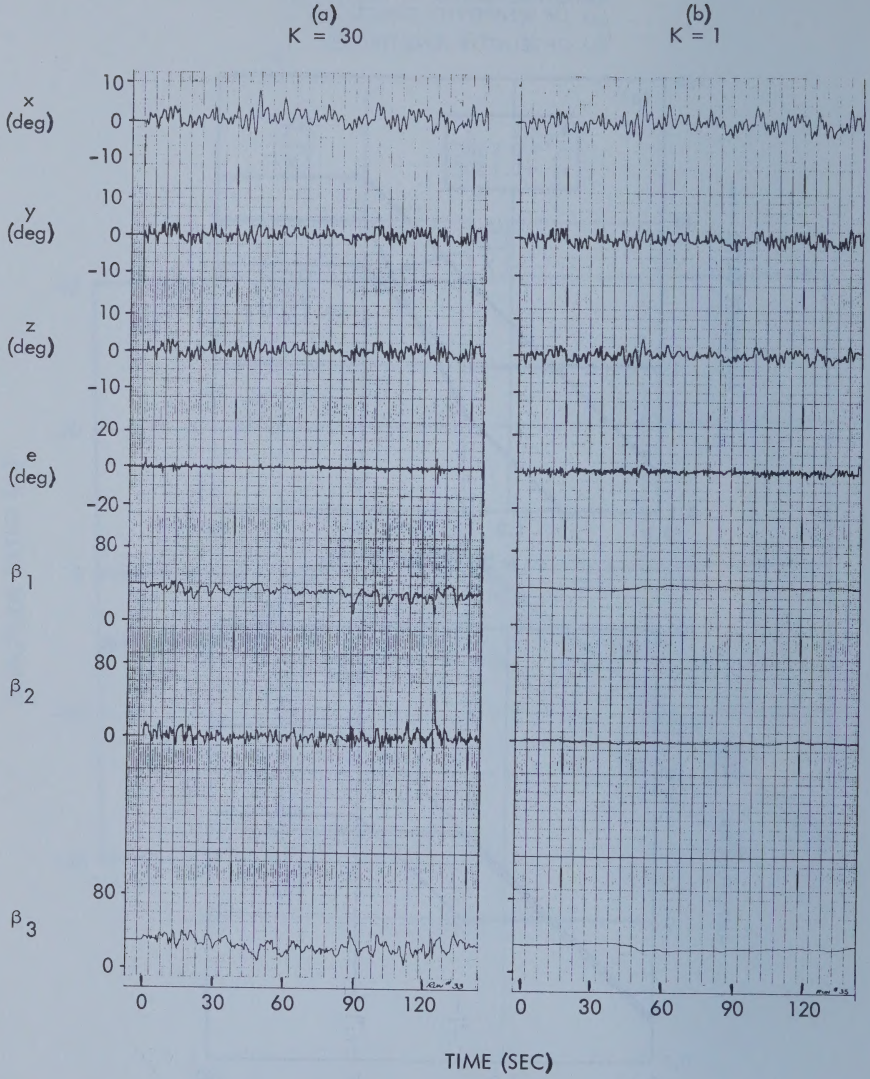


Figure 4. Parameter Time History for a First-Order Model of a Human Control Operator; (a) High Adjustment Gain, (b) Low Adjustment Gain

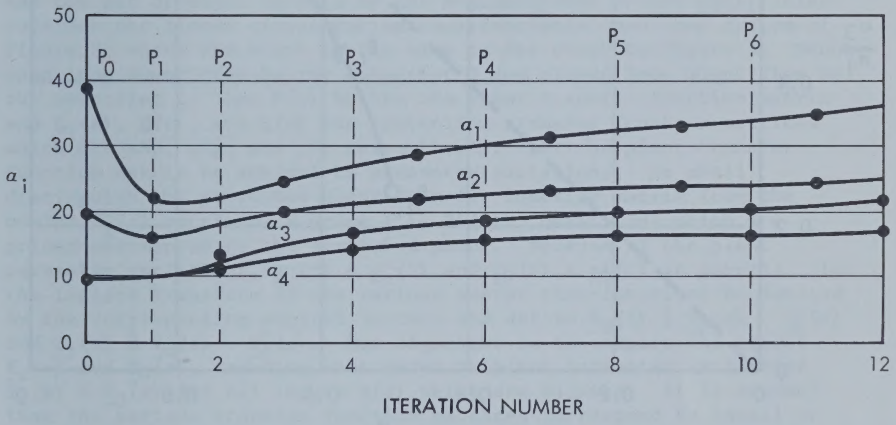


Figure 5. Indeterminate Parameters of a Second-Order Model of a Human Control Operator

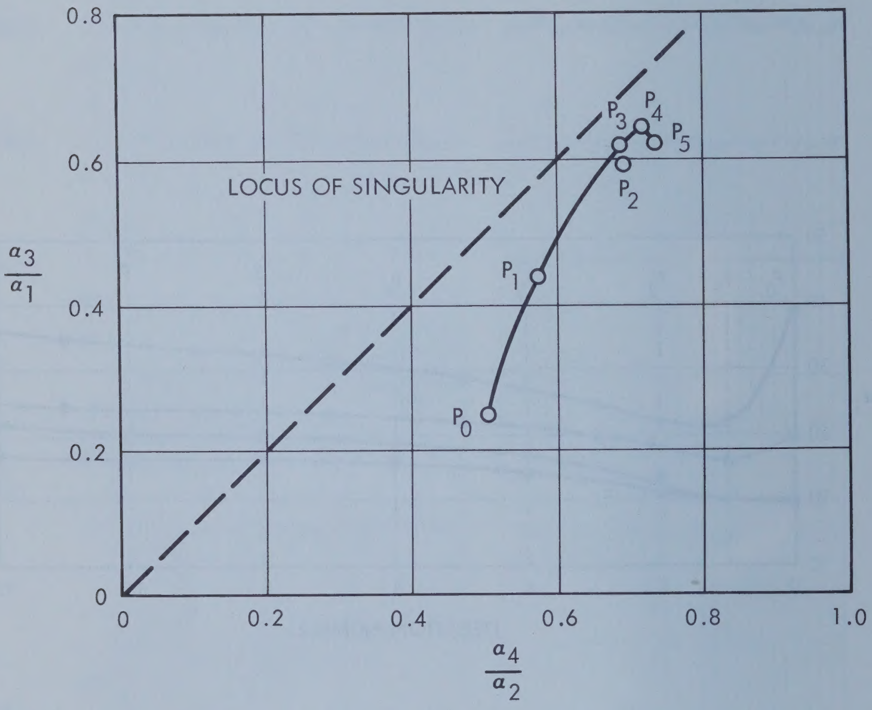


Figure 6. Parameter Ratios $\frac{a_2}{a_4}$ vs. $\frac{a_1}{a_3}$ for Results Shown in Figure 5

SENSITIVITY COMPARISON OF OPEN-LOOP AND CLOSED-LOOP SYSTEMS

J. B. CRUZ, Jr. and W. R. PERKINS
University of Illinois
Urbana, Illinois

INTRODUCTION

The main purpose of this note is to supply a rigorous proof for a frequency domain criterion for sensitivity obtained earlier by the authors [1,2]. First, the results based on the comparison viewpoint are restated. Although analogous results are available for the linear time-varying case [3], only the linear time-invariant but multivariable system is treated for simplicity.

Consider a linear time-invariant multivariable feedback system as in Figure 1, where the plant has m inputs and n outputs. Denote the set of time-domain command input signals by the p-dimensional vector $\underset{\sim}{r}(t)$, the set of inputs to the plant by the m-dimensional vector $\underset{\sim}{u}(t)$ and the set of plant outputs by the n-dimensional vector $\underset{\sim}{y}(t)$. Also consider the linear time-invariant multivariable open loop system of Figure 2, where the plant is the same as the plant in Figure 1. Denote open loop quantities by the subscript $\underline{0}$ and closed loop quantities by the subscript $\underline{c}$. Let $P(s)$ be the $n \times m$ plant transfer function matrix and $\underset{\sim}{G}_1(s)$, $\underset{\sim}{G}(s)$, and $\underset{\sim}{H}(s)$ the controller transfer function matrices which are $m \times p$, $m \times p$, and $p \times n$ respectively. Let the plant transfer function matrix be subject to parameter variations. We shall distinguish the perturbed plant transfer function matrix from the nominal $\underset{\sim}{P}(s)$ matrix by a prime ('), $\underset{\sim}{P}'(s)$. Quantities which are primed correspond to the perturbed plant. Because of the plant parameter variations, $\underset{\sim}{y}_c(t) \neq \underset{\sim}{y}_c'(t)$ and $\underset{\sim}{y}_0(t) \neq \underset{\sim}{y}_0'(t)$ in general. Let the Laplace transform of the various vector time-functions be denoted by the corresponding capital letters and define $\underset{\sim}{E}_0(s) \triangleq \underset{\sim}{Y}_0(s) - \underset{\sim}{Y}_0'(s)$ and $\underset{\sim}{E}_c(s) \triangleq \underset{\sim}{Y}_c(s) - \underset{\sim}{Y}_c'(s)$. Our objective is to compare the errors $\underset{\sim}{E}_c(s)$ and $\underset{\sim}{E}_0(s)$, assuming that under no plant parameter variations, $\underset{\sim}{Y}_c(s) = \underset{\sim}{Y}_0(s)$ for all inputs $\underset{\sim}{R}(s)$ which are allowed. It is assumed that the various transfer function matrices correspond to causal or physically realizable systems. Finally, we assume that the feedback system is realizable in the sense that a transfer function matrix from $\underset{\sim}{R}(s)$ to $\underset{\sim}{Y}(s)$ exists so that any $\underset{\sim}{R}(s)$ produces a unique $\underset{\sim}{Y}(s)$, for all allowable parameter variations.

The first result is that $\underset{\sim}{E}_c(s)$ is linearly related to $\underset{\sim}{E}_0(s)$, i.e.,

$$\underset{\sim}{E}_c(s) = \underset{\sim}{S}(s)\ \underset{\sim}{E}_0(s) \qquad (1)$$

and that

$$\underset{\sim}{S}(s) = [\underset{\sim}{I}_n + \underset{\sim}{P}'(s)\underset{\sim}{G}(s)\underset{\sim}{H}(s)]^{-1} \qquad (2)$$

The existence of the indicated inverse in Eq. (2) is guaranteed by the assumption of existence of a transfer function from $\underset{\sim}{R}(s)$ to $\underset{\sim}{Y}(s)$. This result was derived in references 1 and 2. The matrix $\underset{\sim}{S}(s)$ is defined as the sensitivity matrix for the multivariable system. It may be interpreted as a transfer function matrix for a system whose output is $\underset{\sim}{E}_c$ when the input is $\underset{\sim}{E}_0$. This transfer function matrix may be obtained from the original system by inserting a vector adder at the output of the plant in Figure 1, where the inputs to the adder are $\underset{\sim}{E}_0$ and the plant output. If $\underset{\sim}{R}$ is set to zero, then the system output is easily verified to be $\underset{\sim}{E}_c$. The relationship between $\underset{\sim}{E}_c$ and $\underset{\sim}{E}_0$ can be utilized to directly compare the sensitivities of open and closed loop systems.

As a measure for the error we may use

$$I \triangleq \int_0^{t_1} \underset{\sim}{e}^T(t)\,\underset{\sim}{e}(t)\,dt \tag{3}$$

and we may compare two systems with the same plant, system 1 and system 2, by comparing I_1 and I_2. In particular we may compare the closed loop index I_c to the open loop index I_0. A closed loop system is said to be less sensitive than an open loop system if $I_c \leq I_0$ for all allowed inputs $\underset{\sim}{r}(t)$ and for all t_1. The open loop and closed loop systems have identical transfer function matrices relating $\underset{\sim}{R}(s)$ to $\underset{\sim}{Y}(s)$, whenever there is no plant parameter variation. For the same perturbed plant $\underset{\sim}{P}'(s)$ in the two systems, if $I_c \leq I_0$ no matter what the input is (in the class of allowed inputs) and no matter what t_1 is, then the closed loop system is less sensitive than the open loop system. The class of allowed inputs $r(t)$ which we consider is the class for which $\int_0^{t_1} \underset{\sim}{r}^T(t)\underset{\sim}{r}(t)\,dt < \infty$ for any t_1.

We assume that $\underset{\sim}{S}$ is a stable transfer function matrix in the sense that if the energy of the input I_0 is bounded for $t_1 = \infty$, the output energy I_c is bounded also, for $t_1 = \infty$. Furthermore we assume that whenever I_0 is bounded, so is I_c for any finite t_1. Then, in order for the closed loop system to be less sensitive than the corresponding open loop system, it is sufficient that

$$\underset{\sim}{I}_n - \underset{\sim}{S}^*(j\omega)\,\underset{\sim}{S}(j\omega) \geq 0 \; . \tag{4}$$

be positive semi-definite for all real ω, where $\underset{\sim}{S}^*(j\omega)$ denotes the conjugate transpose of $\underset{\sim}{S}(j\omega)$ and I_n in the n-th order identity matrix. Furthermore, if the transfer function matrix from $\underset{\sim}{R}$ to $\underset{\sim}{E}_0$ is non-singular, then (4) is a necessary and sufficient condition for the closed loop system to be less sensitive than the open loop system. The proof for the validity of this criterion follows immediately from the Theorem in the next section.

In concluding this section, we note that the comparison sensitivity matrix of Eq. (1) places in evidence the connection between the open loop error vector and the closed loop error vector, for any input $\underset{\sim}{r}(t)$ in the class of allowed inputs. That is, S is the matrix that relates the output error for a system with no feedback to the output error for a system with feedback. This definition has a completely different physical basis than the reciprocal of Bode's definition which is a logarithmic change of transfer function with respect to a logarithmic change in a parameter. Moreover, our definition applies to multivariable as well as to single variable systems. But if the system is indeed single variable, and if the parameter variations are differentially small, then our definition yields the same expression as that of the reciprocal of Bode's. Our definition has the additional advantages of a natural physical interpretation, applicability to multivariable systems, multiple plant parameter variations, and large parameter variations. A second comment is that the positive semi-definiteness criterion in Eq. (4) reduces to the familiar $|S(j\omega)| \leq 1$ for the scalar case. Thus for the scalar case, the classical Bode criterion acquires new significance in the sense of reducing the integrated square error of the output.

CRITERION FOR NON-INCREASE OF FINITE-TIME ENERGY OF SIGNALS MAPPED BY A LINEAR TIME-INVARIANT CAUSAL OPERATOR

In this section we state precisely the necessary and sufficient conditions in the frequency domain, for the finite-time energy of the closed loop error to be at most equal to the finite-time energy of the open loop error.

Let $\mathcal{H}_N(0,\infty)$ denote the set of real measurable N-vector-valued functions of the real variable t defined in $[0,\infty)$. Denote the transpose, and the complex-conjugate transpose of a matrix A by A^T and A^*, respectively. Let $\underset{\sim}{1}_M$ denote the identity matrix of order M. Let

$$\mathcal{L}_{2N}(0,\infty) = \{\underset{\sim}{f} \,|\, \underset{\sim}{f} \in \mathcal{H}_N(0,\infty),\ \int_0^\infty \underset{\sim}{f}^T(t)\,\underset{\sim}{f}(t)\,dt < \infty\}. \qquad (5)$$

Let $y \in (0,\infty)$ and define $\underset{\sim}{f}_y$ by

$$\begin{aligned} \underset{\sim}{f}_y(t) &= \underset{\sim}{f}(t) && \text{for } t \in [0,y] \\ &= \underset{\sim}{0} && \text{for } t > y \end{aligned} \qquad (6)$$

for any $\underset{\sim}{f} \in \mathcal{H}_N(0,\infty)$. Let

$$\mathcal{E}_N = \{\underset{\sim}{f} \,|\, \underset{\sim}{f} \in \mathcal{H}_N(0,\infty),\ \underset{\sim}{f}_y \in \mathcal{L}_{2N}(0,\infty) \text{ for all } y \in (0,\infty)\}. \qquad (7)$$

A mapping $\underline{S}$ of $\mathcal{E}_N$ into itself is said to be "causal" if and only if $(\underline{S}\underset{\sim}{f})_y = (\underline{S}\underset{\sim}{f}_y)_y$ for all $\underset{\sim}{f} \in \mathcal{E}_N$ and all $y \in (0,\infty)$. Finally, let $\underset{\sim}{Q}$ denote a real constant m×n "weighting" matrix, and let $\|\cdot\|$ be defined by

$$\|\underset{\sim}{f}\| = \int_0^\infty [\underset{\approx}{Q}\underset{\sim}{f}(t)]^T[\underset{\approx}{Q}\underset{\sim}{f}(t)]dt \text{ for all } \underset{\sim}{f} \in \mathcal{L}_{2N}(0,\infty). \qquad (8)$$

Our result is as follows.

Proposition: Let $\underline{S}$ denote a causal linear time-invariant mapping of $\mathcal{E}_N$ into itself. Let $\underline{S}$ map $\mathcal{L}_{2N}(0,\infty)$ into itself. Let $\underline{S}$ be defined by the condition that if $\underset{\sim}{f} \in \mathcal{L}_{2N}(0,\infty)$ and $\underset{\sim}{g} = \underline{S}\underset{\sim}{f}$, then $\underset{\sim}{G}(j\omega) = \underset{\sim}{S}(j\omega)\underset{\sim}{F}(j\omega)$, for almost all $\omega \in (-\infty,\infty)$ in which $\underset{\sim}{S}(j\omega)$ is an n×n matrix-valued function with essentially bounded elements and $\underset{\sim}{G}(j\omega)$ and $\underset{\sim}{F}(j\omega)$ are the Fourier transforms of $\underset{\sim}{f}$ and $\underset{\sim}{g}$, respectively. Then the following two statements are equivalent.

(i) $\underset{\sim}{Q}^T\underset{\sim}{Q} - \underset{\sim}{S}^*(j\omega)\underset{\sim}{Q}^T\underset{\sim}{Q}\underset{\sim}{S}(j\omega)$ is nonnegative definite for almost all $\omega \in (0,\infty)$

(ii) $\|\underset{\sim}{g}_y\| \geq \|(\underline{S}\underset{\sim}{g})_y\|$ for all $\underset{\sim}{g} \in \mathcal{E}_N$ and all $y \in (0,\infty)$.

The proposition is an immediate consequence of Parseval's identity and the following

Observation: Let $\underline{S}$ be a (not necessarily linear or time-invariant) causal mapping of $\mathcal{E}_N$ into itself such that $\underline{S}$ maps $\mathcal{L}_{2N}(0,\infty)$ into itself. Then the following two statements are equivalent:

1) $\|\underset{\sim}{f}\| \geq \|\underline{S}\underset{\sim}{f}\|$ for all $\underset{\sim}{f} \in \mathcal{L}_{2N}(0,\infty)$

2) $\|\underset{\sim}{g}_y\| \geq \|(\underline{S}\underset{\sim}{g})_y\|$ for all $\underset{\sim}{g} \in \mathcal{E}_N$ and all $y \in (0,\infty)$.

Proof:

1) $\Rightarrow$ 2)

Let $\underset{\sim}{g} \in \mathcal{E}_N$. Then

$$\|\underset{\sim}{g}_y\| \geq \|\underline{S}\underset{\sim}{g}_y\| \geq \|(\underline{S}\underset{\sim}{g}_y)_y\| = \|(\underline{S}\underset{\sim}{g})_y\| \quad \text{for all } y \in (0,\infty)$$

2) $\Rightarrow$ 1)

Let $\underset{\sim}{f} \in \mathcal{L}_{2N}(0,\infty)$. Then

$$\|\underset{\sim}{f}_y\| \geq \|(\underline{S}\underset{\sim}{f})_y\| \quad \text{for all } y \in (0,\infty).$$

Let $y \to \infty$. Since $\underset{\sim}{f}$ and $\underline{S}\underset{\sim}{f}$ are in $\mathcal{L}_{2N}(0,\infty)$, then (1) is obtained.

Remarks: Let $\underset{\sim}{r}(t) \in \mathcal{E}_P$ and let $\Delta\underline{T}_0$ be a linear causal mapping of $\mathcal{E}_P$ into $\mathcal{E}_N$ and $\mathcal{L}_{2P}$ into $\mathcal{L}_{2N}$. Then (i) is a sufficient condition for (ii) to hold for all $y > 0$, all $\underset{\sim}{r} \in \mathcal{E}_P$, and $\underset{\sim}{g} = \Delta\underline{T}_0\,\underset{\sim}{r}$. In case $P = N$, $\Delta\underline{T}_0^{-1}$ exists, $\Delta\underline{T}_0^{-1}$ maps $\mathcal{E}_N$ into $\mathcal{E}_N$ and $\mathcal{L}_{2N}$ into $\mathcal{L}_{2N}$, then (i) is also a necessary condition for (ii) to hold for all $y > 0$, all $\underset{\sim}{r} \in \mathcal{E}_N$ and $\underset{\sim}{g} = \Delta\underline{T}_0 r$. In the application to sensitivity of the previous section, $\underset{\sim}{g}$ is the open-loop output error, $\underline{S}\underset{\sim}{g}$ is the closed loop output error, and $\underset{\sim}{r}$ is the input, and $\underset{\sim}{S}(j\omega)$ is the sensitivity matrix. A time-discrete version of the proposition where finite sums and z-transforms are used may be proved by the same method as above. Note that Eq. (4) is obtained from condition (i) by choosing $\underset{\sim}{Q}$ to be the nth order identity matrix.

CONCLUSIONS

In a linear multivariable feedback system subject to plant parameter variations, the closed loop output error is a linear mapping of the open loop output error. Secondly, if the system is also time-invariant, then under suitable restrictions, the condition that the finite-time energy of the closed loop error is at most equal to the finite-time energy of the open loop error is equivalent to a frequency domain condition that involves the sensitivity matrix. In this paper, a detailed proof for a precise statement of the latter result was presented.

Acknowledgment

The authors are very grateful to I. W. Sandberg of the Bell Telephone Laboratories for providing the version of the proof in Section 2 and for many stimulating discussions.

This work was supported in part by the Joint Services Electronics Programs (U. S. Army, U. S. Navy, and U. S. Air Force) under Contract No. DA 28 043 AMC 00073(E). Portions of this work were supported by the Air Force Office of Scientific Research under Contract No. AF-AFOSR-931-65.

References

1. J. B. Cruz, Jr. and W. R. Perkins, "A New Approach to the Sensitivity Problem in Multivariable Feedback System Design," IEEE Trans. on Automatic Control, AC-9, pp. 216-223; July, 1964.

2. W. R. Perkins and J. B. Cruz, Jr., "The Parameter Variation Problem in State Feedback Control Systems," J. of Basic Engineering, (Trans. ASME, Series D), Vol. 87, pp. 120-124; March, 1965.

3. W. R. Perkins and J. B. Cruz, Jr., "Sensitivity Operators for Linear Time-Varying Systems," Proc. IFAC Symp. on Sensitivity Analysis, Dubrovnik, Yugoslavia, 1964, to appear.

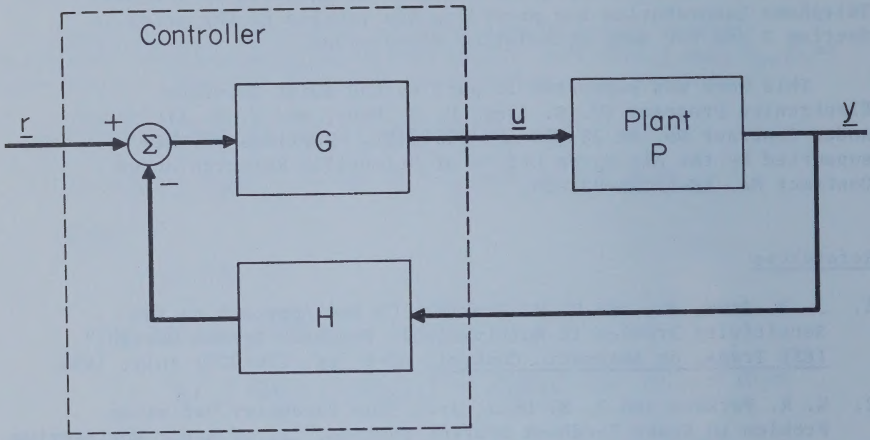

Controller
r
+
Σ
−
G
H
u
Plant
P
y

Figure 1. Linear Time-Invariant Multivariable Closed Loop System.

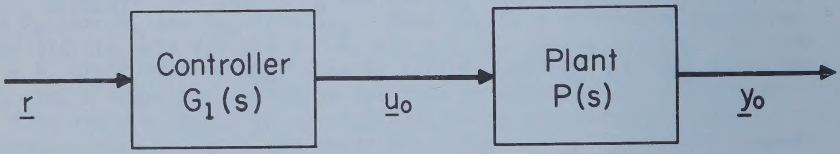

r
Controller
G1(s)
uo
Plant
P(s)
yo

Figure 2. Linear Time-Invariant Multivariable Open Loop System.

SENSITIVITY OF SAMPLED DATA SYSTEMS WITH FINITE SAMPLING WIDTH

L. NARDIZZI
University of Southern California, Electrical Engineering Dept.
Los Angeles, California

ABSTRACT

On the basis of Sensitivity Analysis of discrete systems, a new method is developed for the analysis of sampled-data systems with finite sampling duration. A method for the derivation of the sensitivity function is presented which describes the changes in a parameter dependent upon the pulse width of the sampler. This method can be used to evaluate the response of linear as well as non-linear discrete systems and is illustrated by examples. When particular systems were simulated on an analog computer, the results substantiated the validity of the analysis. Application of the theory to model matching is also discussed briefly.

INTRODUCTION

Sensitivity analysis is now being applied with increasing frequency to sampled data control systems(1, 2, 3). In this respect, this paper represents an application of sensitivity analysis to discrete systems with finite sampling width. The problem of discrete linear systems with finite sampling width has received considerable attention in recent years. The modified z and p transform techniques have been used extensively(4, 5). The use of such transforms is generally limited to linear systems with periodic sampling. A more general approach is the state space concept(6) which allows one to consider non-linear systems and non-periodic sampling. The sensitivity method proposed in this paper is again a different analysis technique which is particularly applicable to computer simulation studies of finite pulsed systems involving non-linear operations and non-periodic sampling.

In this paper the sampling period and the finite pulse width of the sampler are held constant. The finite pulse width of the sampler causes a perturbation of the integral of the error. In other words, the area under the sampled error response (see Figure 1) is the sensitivity parameter and is called the "error-area". This area plays an important role in determining the sensitivity equation of the system. The method applies to systems with finite pulse width samplers of fixed duration. However, the systems can be either open loop or closed loop and may or may not contain hold elements. If no hold element exists in the system to be simulated, a portion of the simulated plant is used as a reconstruction device and the sensitivity is defined in terms of a state variable which is not necessarily the

---

This work was supported in part by the National Aeronautics and Space Administration under Grant NGR 05-018-022.

error. If the sampling is non-periodic, the method applies equally well.

A SENSITIVITY FUNCTION FOR SYSTEMS WITH FINITE PULSE WIDTH SAMPLERS

In the following formulation of the sensitivity function, the closed loop error-sampled data system of Figure 3 is used. The system contains both continuous and discrete elements. The continuous error $e_h(t)$ is sampled by a non-ideal switch which remains closed for a fixed time duration h in each sampling period T. The sampled error $e_h{}^*(t)$ is followed by a zero order hold which feeds into the continuous plant. The output of the zero order hold is approximated by straight line segments in each sampling interval T. The area under this approximated error can be considered to be a perturbation of an error-area of a perfect system (h=0) which is piecewise constant in each sampling interval. The presence of the finite pulse sampler gives rise to the perturbed area which is the parameter to be used in the sensitivity analysis.

The error-area of the non-perfect system of Figure 1 in a sampling period then reduces to

$$\begin{aligned}\overline{A_k} &= \{ Te(t_k) + \frac{h}{2}[e(t_k+h)-e(t_k)] \\ &\quad + [T-h][e_1(t_{k+1}) - e(t_k)]\}\{p(t-t_k) - p(t-t_{k+1})\} \\ &= \{A_k + \Delta A_k\}\{p(t-t_k) - p(t-t_{k+1})\} \qquad (1)\end{aligned}$$

where $e_1(t_{k+1}) = e(t_k+h)$;

$$A_k = Te(t_k) \qquad (2)$$

$$\Delta A_k = [T - \frac{h}{2}][e(t_k+h) - e(t_k)] \qquad (3)$$

and p(t) is the unit step function. The incremental area is approximated by straight lines in the sampling interval. Also note that the finite sampling duration is held constant in each sampling period. The non-homogeneous differential equation describing the system becomes

$$f(y^N, y^{N-1}, \dots y, t) = \frac{\overline{A_k}}{T} \qquad t \in [t_k, t_{k+1}) \qquad (4)$$

such that the right hand side of equation (4) can be considered as a set of parameters. The set can be represented as

$$a = \{ a_0, a_1, \dots , a_k, \dots a_{K-1}\} \qquad (5)$$

where $\quad a_k = A_k \quad$ for $\quad t \in [t_k, t_{k+1})$.

Since each parameter of the set $\{a_k\}$ remains fixed during a sampling interval the sensitivity equation in each period takes on the same form.

$$\frac{\partial^N f}{\partial y^N}\frac{\partial^N y}{\partial a_k^N} + \frac{\partial^{N-1} f}{\partial y^{N-1}}\frac{\partial^{N-1} y}{\partial a_k^{N-1}} + \cdot\cdot \frac{\partial f}{\partial y}\frac{\partial y}{\partial a_k} = \frac{1}{T} \tag{6}$$

The sensitivity function is

$$\mu_k(t) = \frac{\partial y}{\partial a_k}[p(t-t_k) - p(t-t_{k+1})]\, t\varepsilon[t_k, t_{k+1})$$

and $\mu_k^{(n)}(t_k) = 0 \qquad \text{for } n = 1, 2, .. N$ (7)

The sensitivity functions also form a set

$$U = \{\mu_1(t), \mu_2(t), \ldots \mu_k(t)\} \tag{8}$$

The identity of form of equation (6) in each sampling interval provides the solution of one differential equation with the initial conditions (eq. 7) set to zero at the start of each interval. The output of the continuous plant in a sampling interval can now be written as a Taylor series expansion

$$\begin{aligned} \underset{t > t_k}{y(t)} &= y_p(t) + \frac{\partial y_p}{\partial A_k}\Delta A_k + \ldots \\ &\cong y_p(t) + \mu_k(t)\Delta A_k \end{aligned} \tag{9}$$

where $y_p(t)$ represents the response to a perfect sampler (h=0) and $\mu_k(t)\Delta A_k^p$ is the output perturbation.

Example 1. Linear Plant and Zero-order Hold

Consider the system shown in Figure 4. The differential equation describing the continuous portion of the system is given by

$$\ddot{y} = \frac{K(\dot{A} + aA)}{T} \tag{10}$$

where the initial conditions and constants are defined by

$K = 10 \qquad y(o) = 0 \qquad T = .04$ sec.
$a = 14 \qquad y'(o) = 0$

The output of the zero-order hold (perturbed area) in each interval is given by

$$A_k + \Delta A_k = T e(t_k) - \frac{h}{2}[e(t_{k+1}) - e(t_k)] \tag{11}$$

The sensitivity equation in each interval becomes

$$\ddot{\mu}_k = \frac{Ka}{T} \qquad \mu_k(o) = \dot{\mu}_k(o) = 0 \tag{12}$$

The response of this system to a step input for $\frac{h}{T} = .8$ is shown in Figure 7.

Example 2. Model Matching

The example here is intended to show how sensitivity analysis can be used to identify the uniform pulse width h of a sampler. The system in Figure 5 is similar to the previous example. The correction is added to the output of the zero-order hold of the perfect system (h=0) until the output difference between the perfect system and the system to be identified is reduced to zero. The graphs shown in Figure 8 are for $\frac{h}{T} = .6$. The correction added is multiplied by a constant equal to 2 to correct for higher order terms in the Taylor series expansion.

Example 3. Error-pulsed System and Linear Plant

This example is intended to show the validity of the analysis for a system (Figure 6) which does not include a zero-order hold(4). In the analysis a portion of the plant is used to replace the zero-order hold. Figure 2 shows the perfect error area and perturbed error area. The sensitivity equation is calculated with respect to the integral of the variable e_1.

$$\dot{\mu}_k + a\,\mu_k = \frac{K}{T} \qquad \mu_k(o) = \mu_k'(o) = 0 \tag{13}$$

The perturbed area is approximated by a ramp in the interval $0 \le t < h$ and by a step in the interval $h \le t < T$. Therefore the incremental area becomes

$$\Delta A_k = \frac{h}{2}\, e_h(t_k) + \left[T - \frac{h}{2}\right]\left[e_h(t_k + h)\right] - Th\left[e(t_k) + e(t_{k-1})\right] \tag{14}$$

where $$e_h(t_k + h) = e_h(t_k) + h\, e(t_k) \tag{15}$$

substituting eq. (15) into eq. (14) yields

$$\Delta A_k = T\left[\, e_h(t_k) - \frac{h^2}{2T}\, e(t_k) - h\, e(t_{k-1})\right] \tag{16}$$

The results are shown in Figure 9 for a ramp function input.

CONCLUSIONS

Sensitivity analysis can be used to describe sampled data systems with sampling widths of finite fixed duration. The method is equally applicable to systems which are either linear or non-linear and which may or may not contain hold elements. The sensitivity parameter considered in this paper is based on an "error-area". The results of the analysis will depend on the type of approximation used to define the incremental area ΔA. The method was illustrated with examples in which straight line approximations were used for ΔA.

The results illustrate that this method can be easily adapted to hybrid computers in which the error response of a perfect system ($h=0$) can be stored in the digital memory to be used in the calculation of the incremental area.

It should be noted that sensitivity analysis applied to finite pulse systems is a perturbation technique which is computer oriented. If the sampling period is allowed to be non-periodic, the problem involves two perturbed parameters. The combined equations of this non-linear system and the sensitivity equations in general would not be amenable to analytical study.

Another class of promising applications is the study of systems with modulated pulse duration. These problems should result in interesting future work.

REFERENCES

1. Tomovic, R. and Bekey, G.A., "Adaptive Sampling Based on Amplitude Sensitivity", USCEE Report #126, March 1965.
2. Bekey, G.A. and Tomovic, R., "Sensitivity of Discrete Systems to Variation of Sampling Interval", USCEE Report 127, March 1965.
3. Vidal, J., Karplus, W.J. and Keludjian, G., "Sensitivity Coefficients for the Correction of Quantization Errors in Hybrid Computer Systems", Proc. IFAC Symposium on Sensitivity Analysis, Dubrovnik, Yugoslavia, September 1964 (In Press).
4. Farmanfarma, G., "General Analysis and Stability Study of Finite Pulsed Beedback Systems", A.I.E.E. Applications and Industry, July 1958, pp. 148-162.
5. Jury, E.I., Sampled-Data Control Systems, McGraw-Hill Book Company, 1958, pp. 290-377.
6. Tou, J.T., Modern Control Theory, McGraw-Hill Book Company, 1964, pp. 107-110.

ACKNOWLEDGMENT

The author would like to thank Professor Rajko Tomovic for suggesting and encouraging this investigation. The assistance of Mr. Michael Merrit with the analog computer simulation and the guidance of Professor George Bekey are gratefully acknowledged.

LIST OF FIGURES

Figure 1 Output of zero-order hold for system of Figure 4.

Figure 2 Output of reconstruction device $\frac{h}{s}$ for system of Figure 6.

Figure 3 Error sampled discrete data system.

Figure 4 Linear discrete-data control system with adjustment for finite h.

Figure 5 Identification of finite pulse width duration.

Figure 6 Linear discrete data control systems without zero-order hold.

Figure 7 Step function response of system of Figure 4 ($\frac{h}{T} = .8$ $T = .04$ sec.)

Figure 8 Output error response. Model matching examples (see Figure 5, $\frac{h}{T} = .6$, $T = .04$ sec.)

Figure 9 Ramp function response of system of Figure 6 ($\frac{h}{T} = .4$ $T = 1$ sec.)

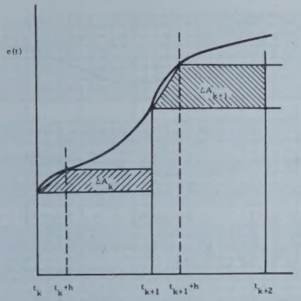


Figure 1

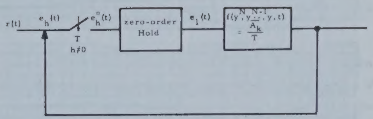

r(t)
$e_h(t)$
$e_h^*(t)$
T
h≠0
zero-order Hold
$e_1(t)$
$f(y^N, y^{N-1}, \ldots, y, t) = \frac{A_k}{T}$

Figure 3

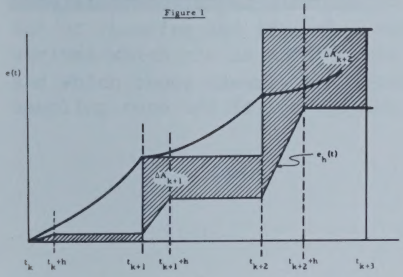


Figure 2

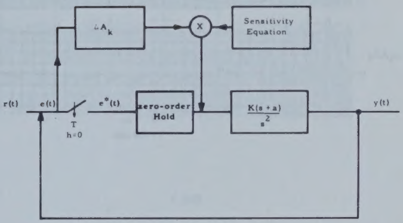

ΔA_k
X
Sensitivity Equation
r(t)
e(t)
T
h=0
$e^*(t)$
zero-order Hold
$\frac{K(s+a)}{s^2}$
y(t)

Figure 4

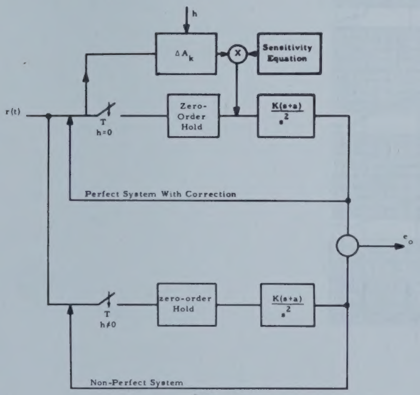

h
ΔA_k
X
Sensitivity Equation
r(t)
T
h=0
Zero-Order Hold
$\frac{K(s+a)}{s^2}$
Perfect System With Correction
e_o
T
h≠0
zero-order Hold
$\frac{K(s+a)}{s^2}$
Non-Perfect System

Figure 5

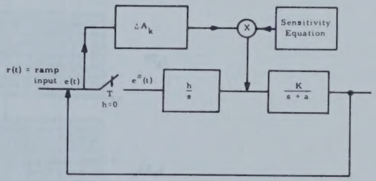

ΔA_k
X
Sensitivity Equation
r(t) = ramp input
e(t)
T
h=0
$e^*(t)$
$\frac{h}{s}$
$\frac{K}{s+a}$

Figure 6

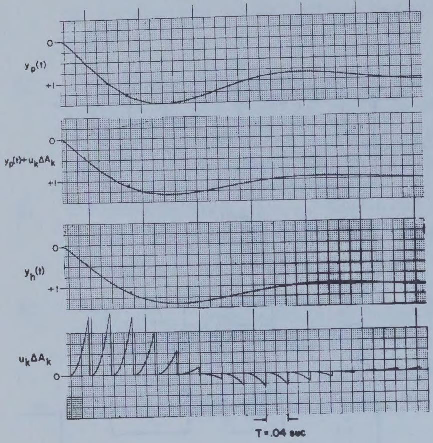


FIG. 7

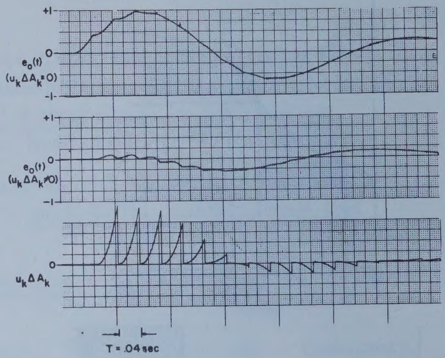


FIG. 8

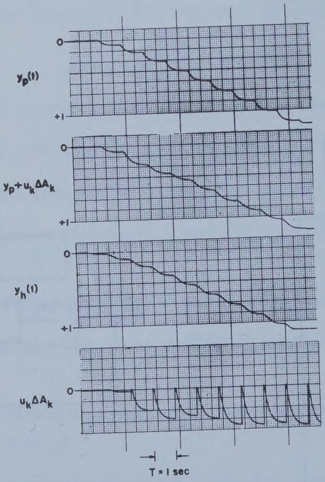


FIG. 9

SOME PRACTICAL CONSIDERATIONS IN THE REALIZATION OF LINEAR DIGITAL FILTERS

J. F. KAISER
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

ABSTRACT

The literature on sampled-data filters, although extensive on design methods, has not treated adequately the important problems connected with the actual realization of the obtained filters with finite arithmetic elements. Beginning with a review of the traditional design procedures a comparison is made between the different canonical realization forms and their related computational procedures. Special attention is directed to the problems of coefficient accuracy and of rounding and trucation effects. A simple expression is derived which yields an estimate of the required coefficient accuracy and which shows clearly the relationship of this accuracy to both sampling rate and filter complexity.

SOME PRACTICAL CONSIDERATIONS IN THE REALIZATION OF LINEAR DIGITAL FILTERS

J. F. KAISER
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

The high speed general purpose digital computer has become a powerful and widely used tool for the simulation of large and complex dynamic systems[1] and for the processing and reducing of large amounts of data by filter methods. The increased computational accuracy of the machines, the broader dynamic ranges in both amplitude and frequency of the system variables, and the increasing order or complexity of the dynamic systems themselves have made it necessary to take a much closer look at the computational and realization details of the designed digital filters. Many of the problems now coming to light were not noticed before[2,3,4,5,6,7] either because the filters were of low order with low (two or three decimal) accuracy requirements or because the sampling frequencies were comparable to the dynamic system and signal frequencies. An understanding of these computational problems and realization considerations is of vital interest to the users of the different digital filter design methods as their presence may often spell the success or failure of a particular application.

The problems to be treated in this paper relate to the numerical determination of the digital filter coefficients, to the stability of the digital filters themselves, to the precision of the arithmetic necessary to carry out the desired filtering operation, and to the choice of filter design methods. The choice of a satisfactory realization scheme is discussed in detail as it pertains to the previously mentioned problems.

Stability and Coefficient Accuracy

The two most widely used methods for the design of digital filters that approximate continuous linear filters with rational transfer characteristics are the bilinear z transformation[8] and the standard z transformation[2] methods.

The bilinear z transform is algebraic in nature and consists simply of a substitution or change of variable in the continuous filter transfer characteristic $H(s)$; i.e., the digital filter $H^*(z^{-1})$ is formed as

$$H^*(z^{-1}) = H(s) \Bigg|_{s \to \frac{2}{T}\frac{(1-z^{-1})}{(1+z^{-1})}} \qquad (1)$$

where

$$z = e^{sT} \qquad (2)$$

the unit advance operator.

The bilinear z transform simply maps the imaginary axis of the s-plane into the unit circle of z^{-1} plane with the left half of the s-plane mapping into the exterior of the unit circle in the z^{-1} plane. The mapping is one-to-one and thus unique. Thus if the transformation indicated by (1) is carried out exactly then $H^*(z^{-1})$ will be stable if $H(s)$ is stable and will be of precisely the same order. The bilinear z form can theoretically be applied directly to a rational transfer characteristic, $H(s)$, in either polynominal or factored form. It will be shown later which form is to be preferred.

The standard z transformation method consists first of obtaining a partial fraction expansion of $H(s)$ in its poles and then z transforming each partial fraction by making use of the transform pair

$$\frac{1}{s+a} \rightarrow \frac{T}{1-e^{-aT}z^{-1}} \tag{3}$$

or transform pairs derived therefrom. The mapping is again such that the left half of the s-plane maps into the exterior of the unit circle in the z^{-1} plane with the portion of the imaginary axis in the s-plane from $n\pi/T$ to $(n+2)\pi/T$, for n any integer, mapping into the unit circle. Thus the standard z transformation when applied to a stable transfer characteristic $H(s)$ always yields an $H^*(z^{-1})$ that is also stable provided the arithmetic is carried out with infinite precision.

For both transforms the resulting $H^*(z^{-1})$ for linear lumped continuous filters is of the general form

$$H^*(z^{-1}) = \frac{\sum_{k=0}^{n} a_k z^{-k}}{1 + \sum_{k=1}^{n} b_k z^{-k}} = \frac{N_d(z^{-1})}{D_d(z^{-1})} = \frac{N_d(z^{-1})}{\prod_{k=1,n} (1-z^{-1}/z_k)} \tag{4}$$

where b_0 has been set unity with no loss in generality. The question now arises as to what accuracy must the coefficients b_k be known to insure that the zeros of $D_d(z^{-1})$ all lie external to the unit circle, the requirement for a stable digital filter. First a crude bound will be established to be followed by a more refined evaluation of the coefficient sensitivity.

The polynomial $D_d(z^{-1})$ can be written in factored form as follows

$$D_d(z^{-1}) = \prod_{k=1}^{n} (1-z^{-1}/z_k) \tag{5}$$

For ease of presentation only simple poles are assumed for the basically low-pass transfer characteristic H(s), there being no difficulty in extending the analysis to the multiple order pole and non low-pass filter cases. If the standard z transform is used then $D_d(z^{-1})$ becomes

$$D_d(z^{-1}) = \prod_{k=1}^{n} (1-e^{p_k T} z^{-1}) \tag{6}$$

where p_k represents the k^{th} pole of H(s) and may be complex. For the bilinear z transform there results

$$D_d(z^{-1}) = \prod_{k=1}^{n} \left[1 - \frac{(1+p_k T/2)}{(1-p_k T/2)} z^{-1} \right] \tag{7}$$

Assuming now that the sampling rate, 1/T, has been chosen to be at least twice the highest critical frequency in the H(s) and in the signals to be processed by the H(s), it is of interest to see how $D_d(z^{-1})$ behaves as the sampling rate is increased further.

The Nyquist limit constraint dictates that for the standard z transformation

$$\mathscr{I}[p_k T] \leq \pi \tag{8}$$

The critical frequencies are normalized with respect to half the sampling frequency as

$$\mu_k = \frac{p_k T}{\pi} = p_k/\omega_n \tag{9}$$

where ω_n is the Nyquist frequency or one-half the sampling frequency. Normally

$$|\mu_k| < 1. \tag{10}$$

Thus as the sampling frequency is increased the μ_k decrease from unity and approach zero. Then one can write for the standard z transform case

$$\left[1 - e^{p_k T} z^{-1}\right] \underset{\text{as } T \to 0}{\Longrightarrow} \left[1 - (1+\mu_k \pi) z^{-1}\right] \tag{11}$$

and for the bilinear case

$$\left[1 - \frac{(1+p_k T/2)}{(1-p_k T/2)} z^{-1}\right] \underset{\text{as } T \to 0}{\Longrightarrow} [1 - (1+\mu_k \pi) z^{-1}] \tag{12}$$

which illustrates that the two design methods yield essentially the same characteristic polynomials, $D_d(z^{-1})$, in the limit as T is made small.

Inspection of (11) and (12) show that the zeros of $D_d(z^{-1})$ tend to cluster about the point $z^{-1} = +1$ in the z^{-1} plane, i.e.,

$$z_k \doteq \frac{1}{1+\mu_k \pi} \approx 1 - \mu_k \pi \tag{13}$$

where for a stable system the μ_k has a negative real part. Now the filter $H^*(z^{-1})$ will become unstable if any of its poles move across the unit circle to the interior as a result of some perturbation or change in the coefficients b_j. To estimate the order of this effect one computes the change necessary to cause a zero of $D_d(z^{-1})$ to occur at the point $z^{-1} = 1$. From (5) and (13) there results

$$D_d(z^{-1}) \Bigg|_{z^{-1}=1} \doteq \prod_{k=1}^{n} (\mu_k \pi) = \prod_{k=1}^{n} (p_k T) \tag{14}$$

But

$$D_d(z^{-1}) \Bigg|_{z^{-1}=1} = 1 + \sum_{k=1}^{n} b_k z^{-k} \Bigg|_{z^{-1}=1} = 1 + \sum_{k=1}^{n} b_k \tag{15}$$

The right hand side of this expression is an important quantity and is therefore defined as

$$F_0 \triangleq 1 + \sum_{k=1}^{n} b_k \tag{16}$$

Thus by combining (14) and (15) it is immediately seen that if any of the b_k are changed by the amount given by (16) then the $D_d(z^{-1})$ will have a zero at $z^{-1} = 1$ and the filter $H^*(z^{-1})$ will thus have a singularity on the stability boundary. A zero of $D_d(z^{-1})$ at $z^{-1} = 1$ causes $H^*(z^{-1})$ to behave as if an integration were present in the $H(s)$. Any further change in the magnitudes of any combination of the b_k in such a manner as to cause $D_d(z^{-1}) \Big|_{z^{-1}=1}$ to change sign will result in an unstable filter, i.e., with some of the zeros of $D_d(z^{-1})$ lying inside the unit circle. Hence (14) is the desired crude bound on coefficient accuracy.

Equation (14) has a significant interpretation; it states that for small μ_k (large sampling rates) the bound on coefficient accuracy is dependent on both the order n of the filter and the sampling rate or normalized filter pole locations. Thus going from an n^{th} order filter to a $(2n)^{th}$ order filter at the same normalized frequency will require approximately twice as many digits accuracy for the representation of the b_k. Similarly doubling the sampling rate for an n^{th} order filter requires $n \times \log_{10} 2$ or $0.3 \times n$ additional decimal digits in the representation of each of the b_k.

Equation (15) has the interpretation that it represents the return difference[9] at zero frequency when the filter $H^*(z^{-1})$ is realized in direct form as shown in Fig. 1. This expression is also recognized as simply the reciprocal of Blackman's deadband factor[10] λ. Thus for complex filters with fairly large sampling rates the quantity F_0 will usually be very small. For example a fifth order Butterworth low pass filter with its break frequency at 1/10 the sampling frequency yields an $H^*(z^{-1})$ having $F_0 = 7.9 \times 10^{-5}$.

The coefficient accuarcy problem is somewhat further aggravated by the fact that as T is made smaller the b_k tend to approach in magnitude the binomial coefficients,[11] $\binom{n}{k}$, and tend to alternate in sign. Thus the evaluation of $D_d^*(z^{-1})$ involves the perennial computational problems associated with the differencing of large numbers. A better bound on coefficient accuracy is obtained by dividing F_0 as obtained from (16) by the magnitude of the largest b_k. The largest b_k is given approximately by

$$\max b_k \approx \binom{n}{[n/2]} \approx \frac{2^n}{\sqrt{n}} \sqrt{\frac{2}{\pi}} \qquad (17)$$

$$\therefore \quad \max b_k \approx \frac{4}{5} * \frac{2^n}{\sqrt{n}}$$

Hence from (14), (16), and (17) an absolute minimum bound on the number of decimal digits m_d required for representing the b_k is found as

$$m_d \geq 1 + \left[- \log_{10}\left(\frac{5\sqrt{n}}{2^{n+2}} \prod_{k=1}^{n} p_k T \right) \right] \tag{18}$$

where [x] denotes the "greatest integer in x."

While the foregoing analysis has yielded an easily computable absolute accuracy bound on the denominator coefficients of the recursive digital filter, the bound is not necessarily the best possible nor does it say anything about what happens to the zeros of $D_d(z^{-1})$ as small perturbations, less than F_0, are introduced in the values of the b_k. It is not enough to say that the digital filter $H^*(z^{-1})$ is simply stable; what is necessary is that the obtained digital filter have response characteristics close in some sense to those of the continuous filter it is approximating. This means that the sensitivity of the zeros of $D_d(z^{-1})$ to changes in the b_k must be determined. The most direct way to establish this relationship is to equate the two forms of the denominator of (4) and then to compute $\partial z_i / \partial b_k$. There results for filters with simple poles only

$$\frac{\partial z_i}{\partial b_k} = \frac{z_i^{k+1}}{\prod\limits_{\substack{k=1 \\ k \neq i}}^{n} \left(1 - \frac{z_i}{z_k} \right)} \tag{19}$$

from which the total differential change in any zero may be evaluated as

$$dz_i = \sum_{k=1}^{n} \frac{\partial z_i}{\partial b_k} \, db_k \tag{20}$$

These results extend directly to the multiple order pole case.[12]

Utilizing the fact that a pole, p_k, of $H(s)$ transforms to a zero of $D_d(z^{-1})$ at $e^{-p_k T}$ for the standard z transform and that for T such that $p_k T \ll 1$ the zero becomes approximately equal to $1 - p_k T$, as given previously by (13) the fractional change δ_i in a zero location z_i can be expressed in terms of the fractional change ε_k in a coefficient b_k of the polynomial $D_d(z^{-1})$. Using (19) there results

$$\varepsilon_k \approx \left[\prod_{\substack{\ell=1 \\ \ell \neq i}}^{n} (z_\ell - z_i) \right] \times \frac{\delta_i}{b_k} \tag{21}$$

where $\varepsilon_k = \dfrac{\Delta b_k}{b_k}$, $\delta_i = \dfrac{\Delta z_i}{z_i}$

and $|1 - z_i| \ll 1$, i.e., tightly clustered zeros have been assumed. For the purpose of quickly estimating the value required for ε_k, the product of the (n-1) factors can be approximated coarsely by the product (14) which is simply F_0. Thus (21) illustrates that the precision required for the representation of the b_k is increased by the factor $\log_{10}(1/\delta_i)$ over that given by (18).

Returning to (19) and (20) it is seen that the detailed changes in the positions of the zeros resulting from changes in the b_k are in general complex functions as the z_i may be complex. The changes in the b_k can occur as a result of imprecise arithmetic used in their computation or as a result of truncating or rounding the obtained b_k coefficients to a smaller number of significant digits. The qualitative evaluation of (19) can also be carried out by using the well developed ideas of the root locus.[13] For example the changes in location of all the zeros of $D_d(z^{-1})$ as a result of a change ρ_k in b_k are found from

$$\left(1 + \sum_{i=1}^{n} b_i z^{-i}\right) + \rho_k z^{-k} = 0 \tag{22}$$

or

$$\prod_{i=1}^{n}\left(1 - \frac{z^{-1}}{z_i}\right) + \rho_k z^{-k} = 0$$

This has the appearance of the standard root locus problem for a single feedback loop having the loop transmission poles of the z_i, a k^{th} order zero at the origin, and a loop gain factor of ρ_k. The parameter F_0 is simply the "gain" ρ_k required when the root locus passes through the point $z^{-1} = 1$. Thus all the techniques of the root locus method and the insight gained thereby can be brought to bear on the problem.

By viewing the coefficient sensitivity problem in terms of root loci the effects of both increasing filter order and especially increasing the sampling rate can be easily observed. Increasing the sampling rate tends to cluster the poles of $H^*(z^{-1})$ even more compactly about the point $z^{-1} = 1$ as Fig. 2 shows for a third order filter. As filter order increases so does the possible order k of the zero at the origin of the z^{-1} plane. All n branches of the root loci begin at the roots z_i; as ρ_k increases k branches converge on the k^{th} order zero at the origin and n-k branches move off toward infinity with eventually radial symmetry. The angles the loci make as they leave the z_i are simply the angles given by evaluating (19) at each z_i. The value of ρ_k at which a branch of the locus first crosses the unit circle (the stability boundary) gives the measure of total variation that can be made in b_k and still keep the filter stable. Clearly the closer the roots z_i are to the unit circle initially the smaller will be the value of ρ_k necessary to move them to lie on the boundary. Thus by varying the ρ_k (the changes in b_k) the extent of the stability problem can be viewed.

The development up to this point assumed that the H(s) was basically low pass with simple poles. Extension to filters of high pass, bandpass, or bandstop types and with multiple order poles presents no real problems when viewed using the root locus idea. For example a digital version of a narrow bandpass filter with center frequency at ω_c would have its poles and zeros located as shown in Fig. 3 about the radial lines at $\pm\omega_c T$ radians. The coefficient sensitivity analysis proceeds in the same way as before except that now the points on the unit circle in the vicinity of $z^{-1} = \cos \omega_c T \pm j \sin \omega_c T$ replace the point $z^{-1} = 1$ in the stability computations. This is easily seen for the standard z transform where a continuous pole of H(s) at $p_k = \sigma_k + j\omega_k$ transforms to a zero of $D_d(z^{-1})$ at $e^{-p_k T}$, i.e., at $e^{-\sigma_k T} \angle \omega_k T$ when written in polar form. Thus it follows that it is primarily the smallness of the real parts of the filter poles of H(s) that cause the z^{-1} plane poles to be very near the unit circle and as a result to contribute measurably to the coefficient accuracy and the related sensitivity problems. It can be shown that expressions quite similar to (14) and (18) can be developed for digital bandpass and bandstop filters. The expressions differ primarily in that for an n^{th} order filter only n/2 terms in the product

$\prod^{n} p_k T$ will in general be small and thus contribute to this measure of sensitivity.

It is interesting to note that in the construction of continuous filters from RLC elements performance is limited primarily by the obtainable Q's of the inductors. The representation of the continuous filter by a digital filter is also strongly influenced by the Q's required of the filter section as it has been shown this directly establishes the number of digits needed to represent the digital filter coefficients.

In this section some of the relationships between filter order, pole locations, sampling frequency and digital filter coefficient accuracy have been established. The question then logically arises, how do these results affect the form chosen for realizing the digital filter? This question is discussed in the following section.

Realization Schemes

The three basic forms for realizing linear digital filters are the direct, the cascade and the parallel forms as shown in Fig. 4. As far as the stability question goes the two variations of the direct form, Fig. 4(a) and Fig. 4(b), are entirely equivalent with the configuration of Fig. 4(a) requiring fewer delay elements. The stability results developed in the previous section indicate clearly that the coefficient accuracy problem will be by far the most acute for the direct form realization. For any reasonably complex filter with steep transitions between pass and stop bands the use of the direct form should be avoided.

The choice between the utilization of either the cascade, Fig. 4(c), or parallel, Fig. 4(d), forms is not clear cut but depends somewhat on the initial form of the continuous filter and the transformation scheme to be used. In any case the denominator of H(s) must be known in factored form. If the parallel form is desired then a partial fraction expansion of H(s) must first be made. This is followed by a direct application of either (1) or (3) if the bilinear or standard z transforms are used respectively. For bandpass or bandstop structures the midfrequency gains of the individual parallel sections may vary considerably in magnitude introducing a small problem of the differencing of large numbers. This parallel form is perhaps the most widely used realization forms.

For cascade realization the bilinear z form requires that the numerator in addition to the denominator of H(s) be known in factored form. The splitting into simpler cascaded forms can then be done rather arbitrarily since the bilinear z operator has the property that

$$Z(G_1G_2) = Z(G_1) \cdot Z(G_2) \qquad (24)$$

If the standard z transform is utilized, a partial fraction expansion must first be made followed by z transforming term by term. Then the fractions must be collapsed to yield an $N_d(z^{-1})$ which must then be factored to permit the cascade realization. This more involved procedure is necessary because the standard z transform does not possess the transform property given by (24).

The discussion up to this point has centered on satisfactory means for obtaining the digital filter coefficients required for the desired realization form. In actually using the digital filter to process data streams the performance of the filter will also be affected by the quantization of the data and by roundoff in the multiplication and addition operations. The recent paper by Knowles and Edwards[14] treats this aspect of the problem in some detail. Their results tend to indicate also that the parallel form realization exhibits slightly less performance error than the cascade form and that the direct form is definitely inferior to both the cascade and parallel forms.

Summary

After reviewing briefly two design procedures for digital filters expressions were derived estimating the accuracy required in the obtained filter coefficients. The expressions showed clearly the relationship of coefficient accuracy to filter complexity and sampling rate. These results also indicated which of the canonical realization forms are to be preferred.

BIBLIOGRAPHY

1. Golden, R. M., "Digital Computer Simulation of Communication Systems Using the Block Diagram Compiler: BLODIB, " Third Annual Allerton Conference on Circuit and System Theory, Monticello, Illinois, October 1965.

2. Ragazzini, J. R. and G. F. Franklin, "Sampled Data Control Systems," McGraw Hill, 1958.

3. Monroe, A. J., "Digital Processes for Sampled Data Systems," John Wiley, 1962.

4. Jury, E. I., "Sampled-Data Control Systems," John Wiley, 1958.

5. Tou, J. T., "Digital and Sampled-Data Control Systems," McGraw Hill, 1959.

6. Jury, E. I., "Theory and Application of the z-Transform Method," John Wiley, 1964.

7. Freeman, H., "Discrete Time Systems," John Wiley, 1965.

8. Kaiser, J. F., "Design Methods for Sampled-Data Filters," Proceedings First Allerton Conference on Circuit and System Theory, November 1963.

9. Bode, H. W., "Network Analysis and Feedback Amplifier Design," Van Nostrand, 1945, pp. 47-49.

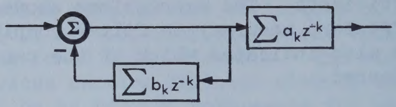
10. Blackman, R. B., "Linear Data-Smoothing and Prediction in Theory and Practice," Addison-Wesley, 1965, p. 76.

11. Mansour, M., "Instability Criteria of Linear Discrete Systems," Automatica Vol. 2, n. 3, January 1965, pp. 167-178.

12. Maley, C. E., "The Effect of Parameters on the Roots of an Equation System," Computer Journal Vol. 4, 1961-2, pp. 62-63.

13. Truxal, J. G., "Automatic Feedback Control System Synthesis," McGraw Hill Book Co., Inc., New York 1955, pp. 223-250.

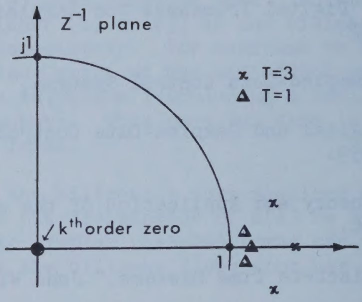
14. Knowles, J. B. and R. Edwards, "Effect of a Finite-Word-Length Computer in a Sampled-Data Feedback System, Proc. IEE Vol. 112, No. 6 June 1965, pp. 1197-1207.



Σ
−
$\sum a_k z^{-k}$
$\sum b_k z^{-k}$

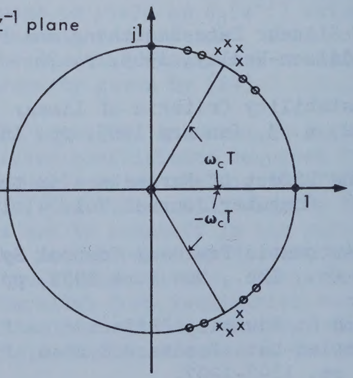
$$H^*(z^{-1}) = \frac{\sum_{0}^{n} a_k z^{-k}}{1 + \sum_{1}^{n} b_k z^{-k}}$$

Fig. 1



Z^{-1} plane
j1
x T=3
Δ T=1
k^{th} order zero
1

Fig. 2



Z^{-1} plane
j1
$\omega_c T$
$-\omega_c T$
1

Fig. 3

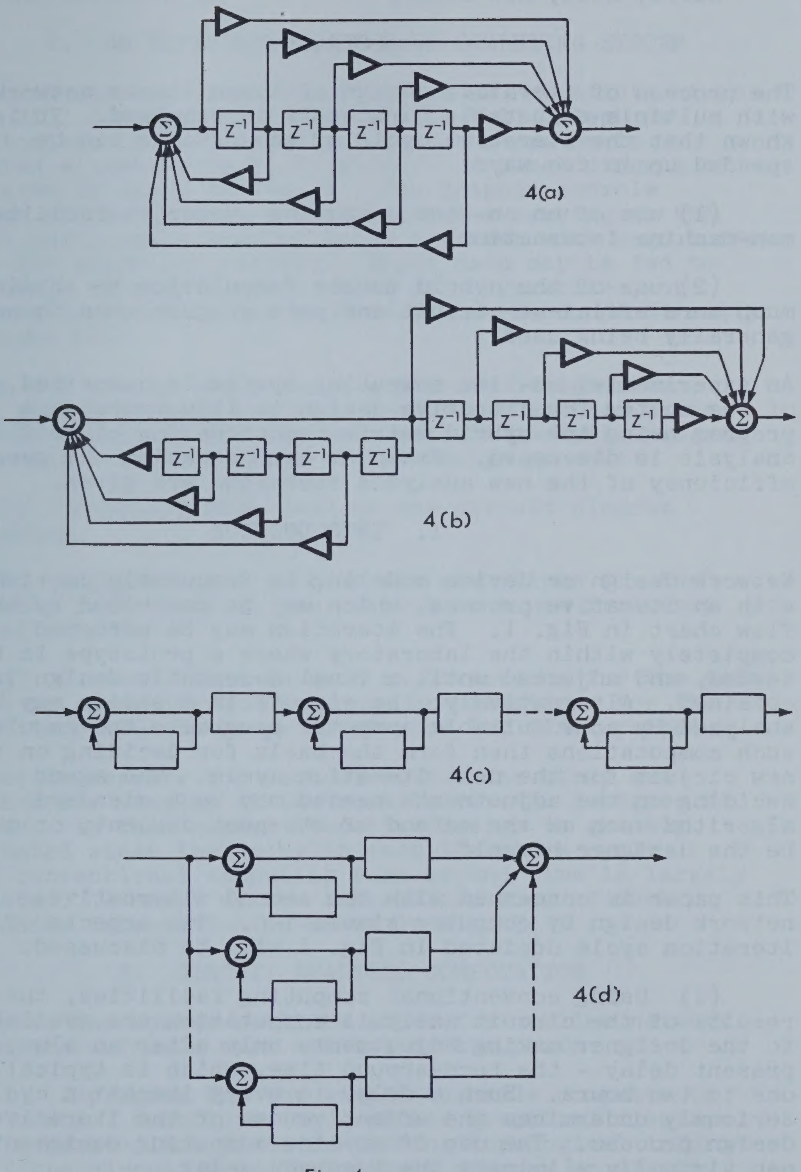

Σ
Z⁻¹
4(a)
4(b)
4(c)
4(d)

Fig. 4

ANALYSIS AND DESIGN OF LINEAR NETWORKS WITH ADJUSTABLE PARAMETERS USING ON-LINE SIMULATION

H. C. SO
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

ABSTRACT

The process of iterative design of large linear networks with multiple adjustable parameters is examined. It is shown that the iterative cycle of the process can be speeded up in two ways:

(1) use of an on-line computing system to facilitate man-machine interaction,

(2) use of the hybrid matrix formulation to obtain a much more efficient circuit analysis program than those generally being used.

An experimental on-line computing system is described. Use of the system for iterative design is illustrated. A program using the hybrid matrix technique for circuit analysis is discussed. Examples demonstrating the greater efficiency of the new analysis technique are given.

1. INTRODUCTION

Network design or device modeling is frequently carried out with an iterative process, which may be described by the flow chart in Fig. 1. The iteration may be effected completely within the laboratory where a prototype is built, tested, and adjusted until a final acceptable design is obtained. Alternatively, the circuit in question may be analyzed by some suitable computer programs; the results of such computations then form the basis for deciding on the new circuit for the next iteration cycle. The agent deciding on the adjustments needed may be a standard algorithm such as the method of steepest descent, or may be the designer himself.

This paper is concerned with the second alternative: network design by computer simulation. Two aspects of the iteration cycle depicted in Fig. 1 will be discussed.

(1) Using conventional computing facilities, the results of the circuit analysis computation are available to the designer making adjustments only after an always-present delay - the turn-around time, which is typically one to two hours. Such a delay in every iteration cycle seriously undermines the effectiveness of the iterative design process. The use of on-line computing equipments can virtually eliminate the feedback delay.

(2) The usual node or loop analysis of circuits requires lengthy computation time, which is particularly objectionable for on-line programs. It is shown that the use of the hybrid matrix formulation[1] to analyze linear networks with adjustable parameters results in much greater computing efficiency.

2. AN EXPERIMENTAL ON-LINE COMPUTING SYSTEM

An experimental on-line computing system now exists at the Murray Hill Computation Center of the Bell Telephone Laboratories. A full description of the system is presented elsewhere by W. H. Ninke.[2] A block diagram of the system is shown in Fig. 2. The graphic console consists of two units: a small computer serving as the control unit, and a display scope together with core buffer memory for graphical display. Input data may be fed to the console through a card reader or a typewriter, or graphical information may be entered on the scope directly by a light pen.

Through the use of the console a network designer can easily communicate to the main computer the three types of inputs necessary for the main circuit analysis program during each iteration cycle of the design process:

(1) Graphical data such as the circuit diagram under study.

(2) Adjusted data such as the new values for the adjustable parameters in the circuit or a new frequency range of interest.

(3) Control commands such as directions for the main program to change frequency range or directions for the form of intermediate scope display or for final output.

Computation results for each circuit under study are sent back for display at the console for the designer to act upon. In this manner iterative design is greatly facilitated since the delay in each iteration cycle due to conventional computing turn-around time is largely eliminated. A chart of the information flow of such a man-machine interaction is shown in Fig. 3.

3. CIRCUIT ANALYSIS COMPUTATION

To support the system outlined above for on-line network design, the software must consist of two parts:

(1) A program for the graphic console computer to effect control functions, to generate graphical displays, and to generate link lists containing topological information of the circuit from the circuit diagram on display.

(2) A program for the main computer. This program in turn must contain two parts: the first part effects control functions and converts commands and data received from the console to forms usable by the second part, which actually carries out the analysis computation.

For both computers in the system, programs are being written for graphical display and data conversion. A discussion of these programs is not relevant here. However, the circuit analysis computation program, which can be used for frequency analysis of large linear networks, is of interest here since it employs a new circuit analysis technique that results in greater computing efficiency.

Basically, if a linear network is given and some frequency characteristic of the network is required, the problem is routine. A set of steady-state simultaneous network equations can be set up and solved for at each frequency point. Thus one may use, for example, the cut-set equations for the two-port in Fig. 4:

$$Y_C \vec{V}_C = \vec{I}_C \qquad (1)$$

where $\vec{V}_C = (v_1, v_2, v_3, \ldots v_N)'$, v's complex numbers,

$\vec{I}_C = (j_1, j_2, 0, \ldots 0)'$, j's complex numbers,

and Y_C is the NxN cut-set admittance matrix of the N+1 node network at a particular frequency f. To obtain the short-circuit admittance matrix Y_{SC} of the two-port at f, Gaussian reduction[3] is applied to Y_C so that

$$Y_C \rightarrow \begin{array}{c} \begin{array}{cc} \overbrace{\phantom{Y_{SC}}}^{2} & \overbrace{}^{N-2} \end{array} \\ \left[\begin{array}{c|c} Y_{SC} & 0 \\ \hline B & D \end{array}\right] \begin{array}{l} \}2 \\ \}N-2 \end{array} \end{array} \qquad (2)$$

where D is the result of triangulating the corresponding (N-2)x(N-2) submatrix of Y_C; that is to say, all entries of D above the diagonal are zero.

If one denotes the time required to form Y_C by T_Y, the time to perform the matrix reduction indicated in (2) by $T_{N\rightarrow 2}$, and if N_f is the number of frequencies of interest, then the time required to compute a frequency characteristic by this commonly used technique is $N_f \times (T_Y + T_{N\rightarrow 2})$. Now if there are r adjustable parameters inside the two-port, and if N_p is the number of different parameter settings, called trials, used in each iteration cycle of the design process, then the total computation time per cycle is given by

$$T_c = N_f \times N_p \times (T_Y + T_{N\rightarrow 2}) \qquad (3)$$

The important point here is that when this conventional method of steady-state frequency analysis of networks is used for problems involving adjustable parameters, the whole process of reducing the NxN system matrix to the 2x2 final matrix has to be repeated not only at each different frequency but also at each different parameter setting (trial).

An alternative approach is the following: pull all r adjustable parameters outside the two-port and consider what remains as a linear, possibly active, m-port, $m \geq r+2$. If all adjustable elements are one-port elements then $m = r+2$. If, for example, a transistor is considered to be an adjustable parameter then pulling such a parameter outside will of course increase the number of ports by two. The m-port is then described by the hybrid matrix equation*

$$H\vec{X} = \vec{W}, \qquad H = \left[\begin{array}{c|c} H_{11} & H_{12} \\ \hline H_{21} & H_{22} \end{array}\right] \tag{4}$$

where $\vec{X} = (x_1, x_2 \mid x_{r_1}, x_{r_2}, \ldots x_{r_{m-2}})' = [\vec{\alpha} \mid \vec{\beta}]'$

is the vector of independent, in general mixed, port

variables, and $\vec{W} = (w_1, w_2 \mid w_{r_1}, w_{r_2}, \ldots w_{r_{m-2}})' = [\vec{\theta} \mid \vec{\Phi}]'$

is the vector of corresponding dependent port variables. In any trial, the parameters located at the m-2 ports determine another matrix equation:

$$\vec{\Phi}' = \rho\vec{\beta}' \tag{5}$$

Combining (4) and (5),

$$H_{AUG}\vec{X} = \left[\begin{array}{c|c} H_{11} & H_{12} \\ \hline H_{21} & (H_{22}-\rho) \end{array}\right] \left[\begin{array}{c} \vec{\alpha}' \\ \hline \vec{\beta}' \end{array}\right] = \left[\begin{array}{c} \vec{\theta}' \\ \hline 0 \end{array}\right] \tag{6}$$

Thus at any frequency and for any trial, the input-output relation of the two-port in Fig. 4 is obtained by applying Gaussian reduction to H_{AUG} so that

$$H_{AUG} \rightarrow \begin{array}{c} \begin{array}{cc} \overbrace{\phantom{H_{TP}}}^{2} & \overbrace{}^{m-2} \end{array} \\ \left[\begin{array}{c|c} H_{TP} & 0 \\ \hline \tau & \delta \end{array}\right] \begin{array}{l} \}2 \\ \}m-2 \end{array} \end{array} \tag{7}$$

---

*Since the adjustable parameters may be distributed in any manner in the original network, the open-circuit impedance matrix or short-circuit admittance matrix of the resulting n-port frequently does not exist. See Reference 1.

where δ is the result of triangulating the corresponding $(m-2)x(m-2)$ submatrix of H_{AUG}; that is to say, all entries of δ above the diagonal are zero.

Now consider the scheme shown in Fig. 5 for computing the required frequency characteristics for different trials. Let T_H be the time required to generate the hybrid matrix from the system immittance matrix at one frequency. Let $T_{m\rightarrow 2}$ be the time required to carry out the reduction indicated in (7). Then the total computation time per cycle of the iterative design process is given by

$$t_c = N_f(T_Y+T_H+N_pT_{m\rightarrow 2}) \tag{8}$$

The explicit formulation of the hybrid matrix in terms of topological network equations has been given in a previous paper.[1] An approximation for T_H, when m is not larger than N, is that T_H is equal to or less than twice the time required to reduce the NxN cut-set matrix Y_C (or loop matrix Z_L) to 2x2. Assuming $T_H = 2 \times T_{N\rightarrow 2}$,

$$t_c \simeq N_f(T_Y+2T_{N\rightarrow 2}+N_pT_{m\rightarrow 2}) \tag{9}$$

The efficiency of the hybrid matrix approach can be seen by a comparison of (3) and (9). For m not larger than N,

$$T_c-t_c \simeq N_f([N_p-1]T_Y+[N_p-2]T_{N\rightarrow 2}-N_pT_{m\rightarrow 2}) \tag{10}$$

Usually, $N_p > 2$, $m << N$, and N is large. Then

$$T_c-t_c \simeq N_f([N_p-2] \times T_{N\rightarrow 2} + [N_p-1]T_Y) \tag{11}$$

The advantage of the hybrid matrix approach stems from the fact that for each new trial only a mxm matrix, not NxN, has to be reduced to 2x2. As the time required to reduce a kxk matrix to 2x2 increases very rapidly with k,[4] the improvement in computation efficiency can be very substantial when $m << N$.

However, when $m \simeq N$, t_c and T_c are of the same order of magnitude. When $m >> N$, H degenerates into primarily a matrix of topological incidence submatrices. In this case few operations are required to obtain H so that T_H of equation (8) is negligible. Thus

$$t_c \simeq N_f(N_p \times T_{m\rightarrow 2} + T_Y) \tag{12}$$

Furthermore, because of the degeneracy of H, $T_{m\to 2} \simeq T_{N\to 2}$, and again $t_c \simeq T_c$. These first-order estimates of relative computing efficiency have been borne out by experience.

4. PROGRAM CAPABILITY AND EXAMPLES

A frequency analysis program using the scheme described above (Fig. 5) has been written for the IBM 7094. The program can handle linear networks with RLC elements and any linear elements described by two-port immittance or hybrid matrices. The network to be dealt with may have up to seventy nodes or branches - a RLC series element is considered to be a single branch, but a two-port element is considered to be equivalent to two branches and four nodes for this purpose. However, the network must not require more than thirty cut-set equations and more than thirty loop equations. Maximum number of adjustable branches is thirty. Since the program is written to support an on-line mode of operation, it has a very flexible structure so that entry to the program can be made at several different points. This last feature makes it possible to carry out computation for any new trial with minimum new data input; namely, only new values for adjusted data are required to be fed in. The designer operating at the console is therefore spared the preparation of cumbersome data inputs once initialized of the job is completed.

For comparison of computing efficiency, the two networks shown in Fig. 6 and Fig. 7 were analyzed first by a conventional node-to-datium analysis program and then by HYBRID, the program described here. Both programs use double-precision arithmetic. These computations were made with conventional over-the-counter runs. For each circuit the analysis carried out was similar to what might be required in one iteration cycle in a steepest descent study. The adjustable parameters were changed one at a time and the required characteristic, the complex ratio E_{OUT}/E_{IN}, was calculated. There are eight nodes and seven adjustable parameters, including one transistor, in the network of Fig. 6; consequently $m = 10$ and $N = 7$. The network of Fig. 7 has sixteen nodes and five adjustable parameters including two transistors; hence $m = 9$ and $n = 15$. For the first case, where $m > N$ and $N_p = 8$, HYBRID and node analysis required about the same computation time: about 0.8 seconds for all trials per frequency point. For the second case, where $m << N$ and $N_p = 6$, HYBRID proved to be far superior: the computation time required for all trials per frequency was 1.65 seconds for HYBRID and 5.40 seconds for node analysis.

5. ACKNOWLEDGMENT

The author wishes to thank the following colleagues at the Bell Telephone Laboratories at Murray Hill, New Jersey: E. J. Sitar and Miss D. L. Parker, for their programming support; W. H. Ninke and C. Christensen, for valuable discussions on the experimental graphic console; H. S. McDonald and J. F. Kaiser, for their kind encouragement.

BIBLIOGRAPHY

1. So, H. C., "On the Hybrid Description of a Linear n-Port Resulting from the Extraction of Arbitrarily Specified Elements," IEEE *Trans. on Circuit Theory*, Vol. CT-12, September 1965.

2. Ninke, W. H., "Graphic 1 - A Remote Graphical Display Console System," Fall Joint Computer Conference, November 1965.

3. Faddeev, D. K. and Faddeeva, V. N., *Computational Methods of Linear Algebra*, San Francisco and London: W. H. Freeman and Company, 1963, pp. 128-137.

4. Bodewig, E., *Matrix Calculus*, Amsterdam: North-Holland Publishing Company, 1959, pp. 110-116.

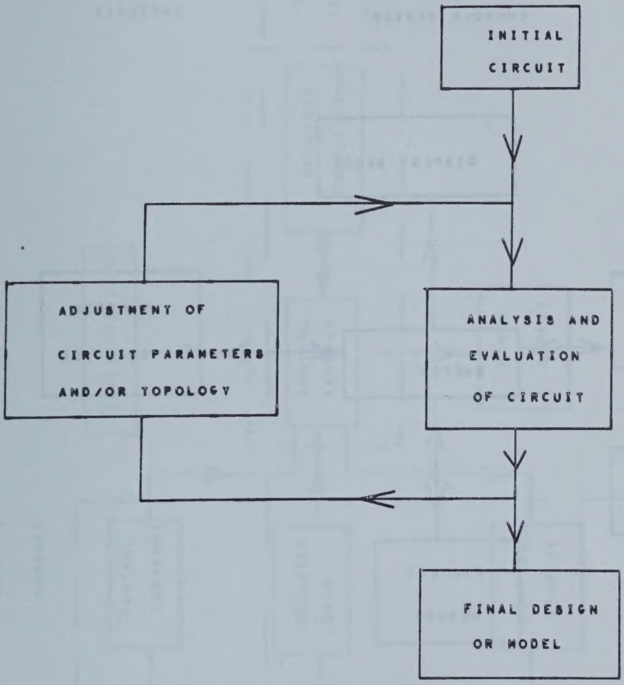

INITIAL CIRCUIT
ADJUSTMENT OF CIRCUIT PARAMETERS AND/OR TOPOLOGY
ANALYSIS AND EVALUATION OF CIRCUIT
FINAL DESIGN OR MODEL

FIG. 1

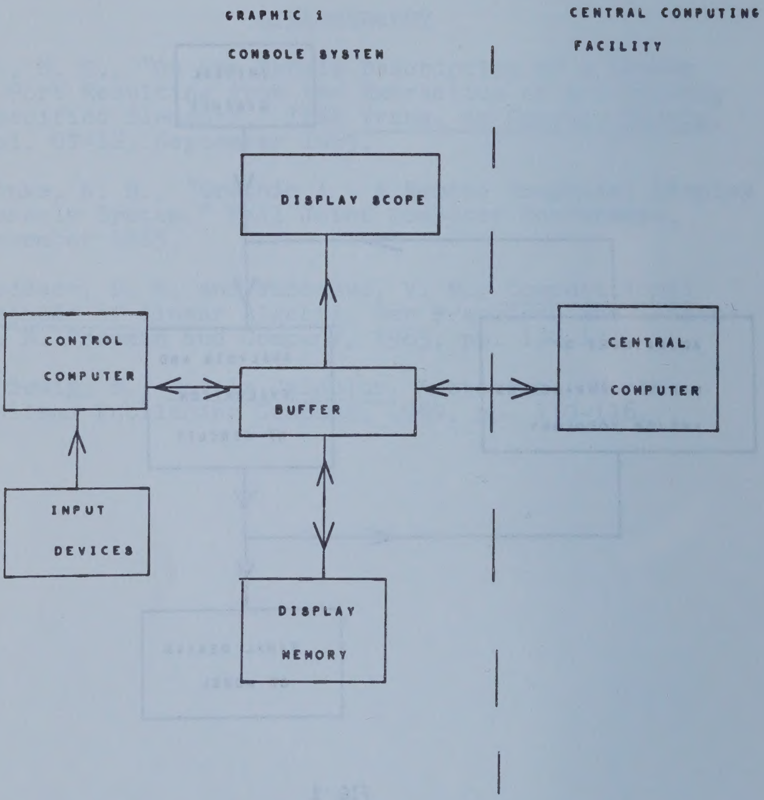

GRAPHIC 1
CONSOLE SYSTEM
CENTRAL COMPUTING
FACILITY
DISPLAY SCOPE
CONTROL
COMPUTER
BUFFER
CENTRAL
COMPUTER
INPUT
DEVICES
DISPLAY
MEMORY

FIG. 2

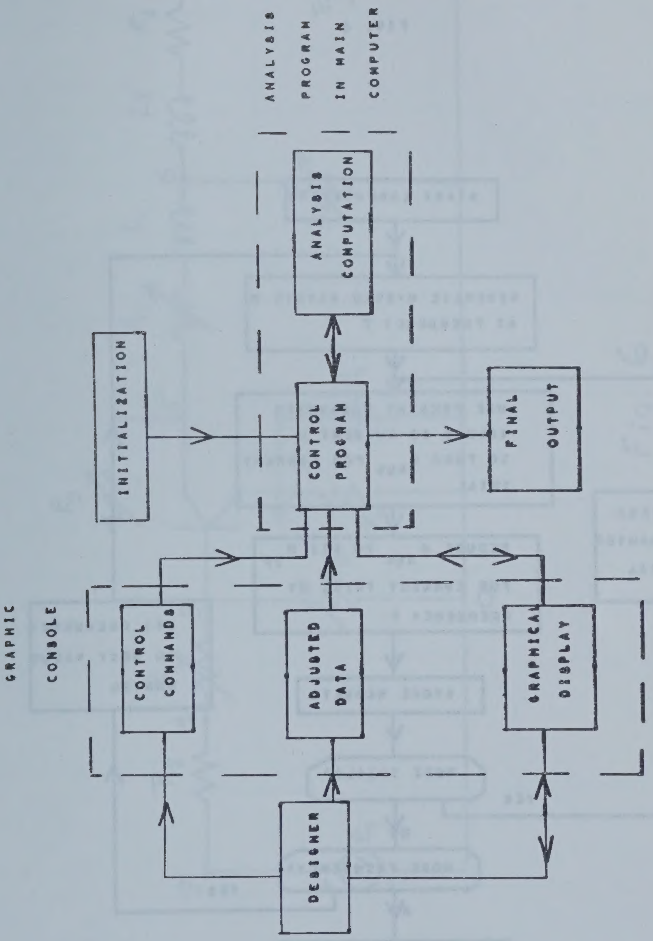

GRAPHIC CONSOLE
CONTROL COMMANDS
ADJUSTED DATA
GRAPHICAL DISPLAY
DESIGNER
INITIALIZATION
CONTROL PROGRAM
ANALYSIS COMPUTATION
FINAL OUTPUT
ANALYSIS PROGRAM IN MAIN COMPUTER

FIG. 3

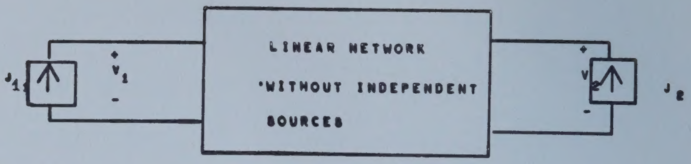

J_1
$+$
V_1
$-$
LINEAR NETWORK
WITHOUT INDEPENDENT
SOURCES
$+$
V_2
$-$
J_2

FIG. 4

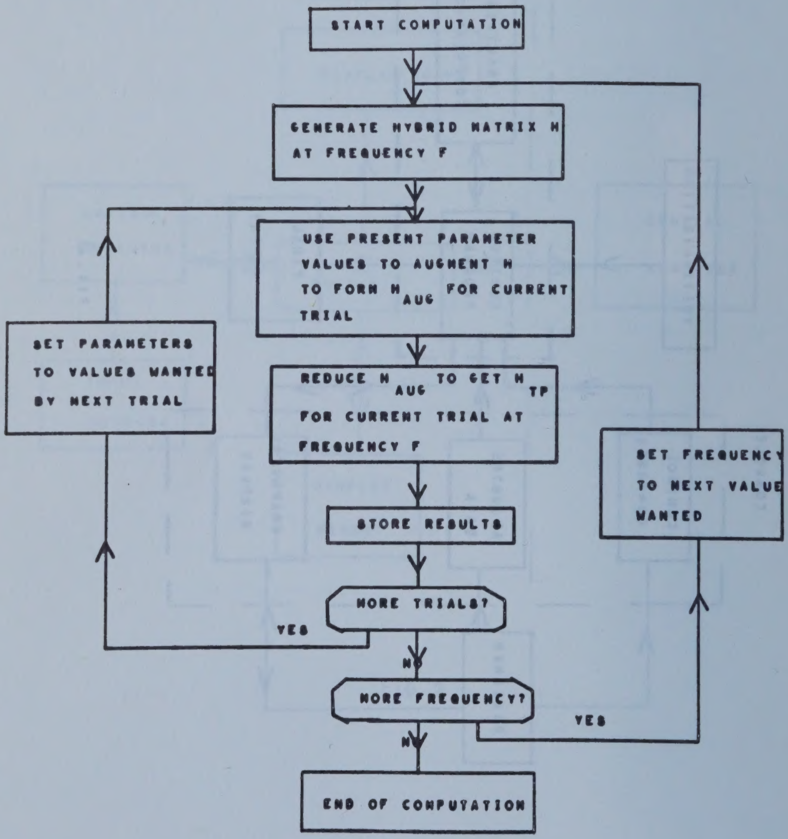

START COMPUTATION
GENERATE HYBRID MATRIX H AT FREQUENCY F
USE PRESENT PARAMETER VALUES TO AUGMENT H TO FORM H_{AUG} FOR CURRENT TRIAL
REDUCE H_{AUG} TO GET H_{TP} FOR CURRENT TRIAL AT FREQUENCY F
STORE RESULTS
MORE TRIALS?
YES
NO
SET PARAMETERS TO VALUES WANTED BY NEXT TRIAL
MORE FREQUENCY?
YES
NO
SET FREQUENCY TO NEXT VALUE WANTED
END OF COMPUTATION

FIG. 5

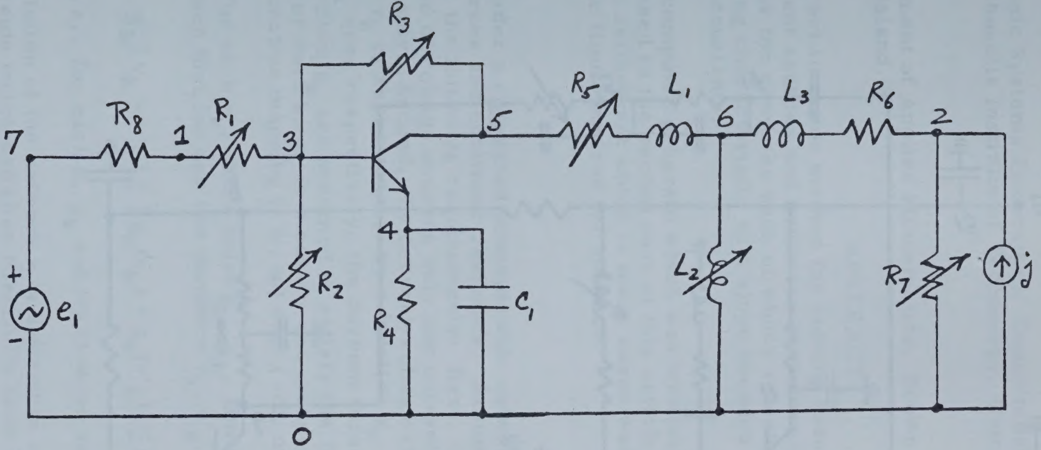

R3
R5
L1
L3
R6
7
R8
1
R1
3
5
6
2
4
+
e_1
-
R2
R4
C1
L2
R7
j
0

Fig. 6

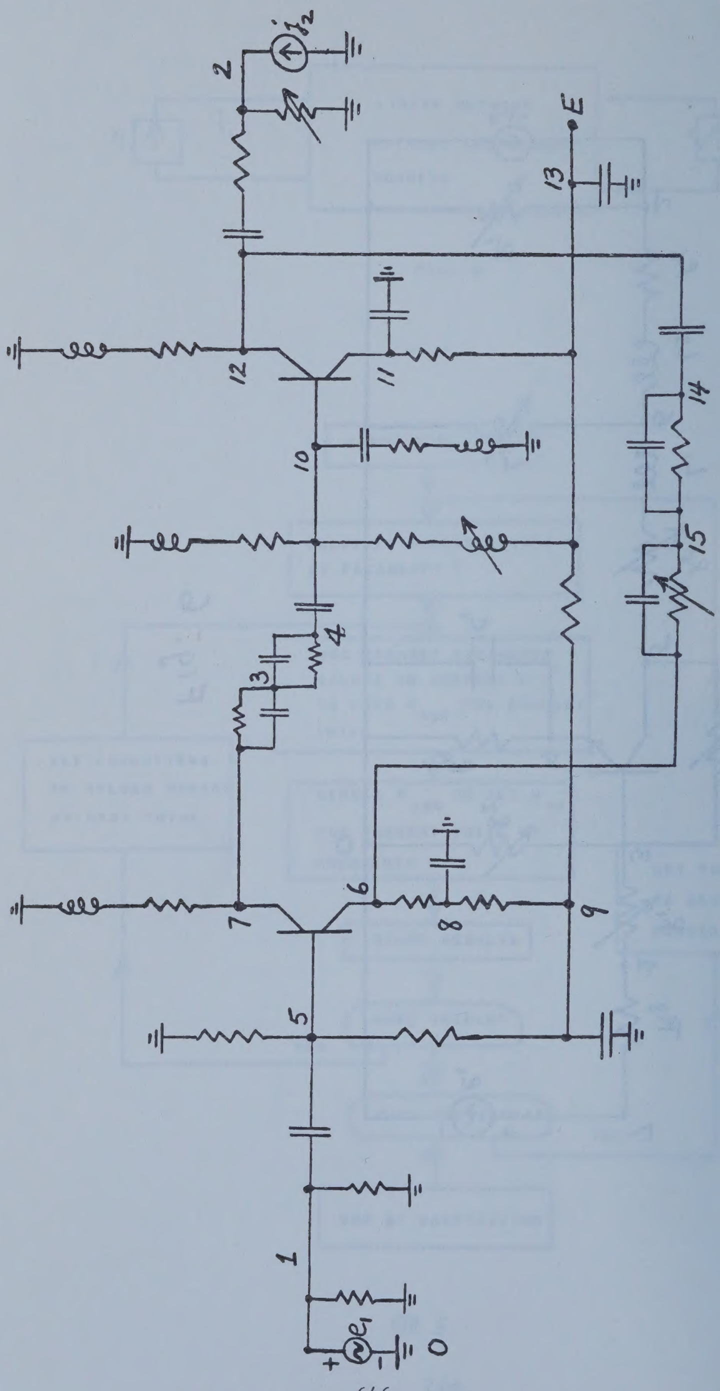

+
-
e1
0
1
5
7
6
8
9
3
4
10
11
12
2
i2
13
E
14
15

Fig. 7

AN ITERATIVE METHOD FOR SOLUTION OF NONLINEAR RESISTOR NETWORKS[†]

JACOB KATZENELSON*   LEON H. SEITELMAN**

* Electronic Systems Laboratory, Department of Electrical Engineering
Massachusetts Institute of Technology, Cambridge, Massachusetts
02139.

** Department of Applied Mathematics, Brown University, Providence,
Rhode Island.

ABSTRACT

A direct iteration method for solving networks composed of independent sources and nonlinear resistors is developed. The method converges for networks each of whose resistors has a monotonic non-decreasing characteristic with slope bounded from above and below by positive constants.

The computer program which was written to implement this method is discussed in the second part of this article. Mention is made of a heuristic refinement which in many cases was found to reduce significantly the time required for solution.

Consider a connected nonseparable network η composed of independent sources and nonlinear resistors. Assume further that the network satisfies the following requirements: first, there are neither loops composed of voltage sources only nor cut-sets composed of current sources only. Second, the current-voltage characteristic of each resistor r_k is representable by a function g_k, with $i_k = g_k(v_k)$, where i_k and v_k are, respectively, the current through and voltage across r_k. The functions g_k are assumed to satisfy the following conditions:

(a) For each k, g_k is a one-to-one monotonic non-decreasing function mapping $(-\infty, \infty)$ onto $(-\infty, \infty)$.

(b) For each g_k, there exist $\overline{g}_k$, $\underline{g}_k$, finite numbers, with $\underline{g}_k > 0$ such that for any two numbers v_k, v'_k,

$$\underline{g}_k | v_k - v'_k | \le | g_k(v_k) - g_k(v'_k) | \le \overline{g}_k | v_k - v'_k |$$

i.e., for each k, g_k and its inverse satisfy Lipschitz conditions.

A solution of the network η is defined to be any set of branch voltages and currents which satisfies Kirchoff's laws and the resistor characteristics. It follows from the work of R. J. Duffin [1] that corresponding to any choice of the source values the network η has a unique solution. It is the purpose of this article to develop an iterative method for the calculation of this solution.

We shall begin by considering the circuit equations; our conventions parallel those in [2].

[†] Work was done while both authors were with Bell Telephone Laboratories, Murray Hill, New Jersey.

Let τ be a tree of the network containing all the voltage sources and no current sources. τ exists; see [2]. Branches which are included in the tree are called tree branches and branches not in the tree are called links.

Let the m voltage source values be labeled $E_1, \ldots, E_m$, the n resistive tree branches $r_1, \ldots, r_n$, the p resistive links $r_{n+1}, \ldots, r_{n+p}$, and the q current sources values $J_1, \ldots, J_q$. Define (where ' denotes transpose)

$$\underline{e} = (e_1, \ldots, e_n)' \quad \text{vector of resistive tree branch voltage}$$

$$\underline{E} = (E_1, \ldots, E_m)' \quad \text{vector of voltage source values}$$

$$\underline{J} = (J_1, \ldots, J_q)' \quad \text{vector of current source values}$$

$$\hat{\underline{e}} = \begin{bmatrix} \underline{E} \\ \\ \underline{e} \end{bmatrix} \quad \text{vector of tree branch voltages}$$

$$\underline{v} = (v_1, \ldots, v_{n+p}) \quad \text{vector of resistive branch voltages}$$

$$\underline{i} = (i_1, \ldots, i_{n+p}) \quad \text{vector of resistive branch currents}$$

Let $\underline{Q} = (q_{ij})$ be the fundamental cut-set matrix of the network η, where the rows of Q are ordered so that they correspond first to voltage sources and last to resistive tree branches, and the columns of $\underline{Q}$ correspond first to voltage sources, then to resistive tree branches, and last to current sources.

Let us make the following notations: We denote by E, R, and J the subsets of the network branches which are voltage sources, resistors, and current sources respectively. T denotes tree branches and L denotes links. T(R) denotes the subset of resistive tree branches while L(R) denotes the subset of resistive links, etc. T(RE) will denote the union of T(R) and T(E) which actually contains all the tree branches.

If X is a subset of the branches of τ, and Y is a subset of the branches of η, then we define $\underline{\gamma}_{X,Y}$ to be that submatrix of $\underline{Q}$ consisting of the intersection of the rows of $\underline{Q}$ corresponding to the elements of X with the columns of $\underline{Q}$ corresponding to the elements of Y.

Thus, $\underline{\gamma}_{T(ER), L(R)}$ denotes the $(m+n) \times p$ submatrix of $\underline{Q}$ which consists of the intersection of the rows corresponding to voltage sources and resistive tree branches with the columns corresponding to the resistive links. The equation of the network is

$$\underline{\gamma}_{T(R),R}\, \underline{\mathcal{G}}^{*}\!\left(\underline{\gamma}'_{T(ER),R} \begin{bmatrix} \underline{E} \\ \underline{e} \end{bmatrix}\right) + \underline{\gamma}_{T(R),L(J)}\underline{J} = \underline{0}, \tag{1}$$

where

$$\underline{\mathcal{G}} = \begin{bmatrix} g_1 & & & 0 \\ & g_2 & & \\ & & \ddots & \\ 0 & & & g_{n+p} \end{bmatrix}$$

is the diagonal matrix of operators, i.e.,

$$\underline{i} = \mathcal{G}^*(\underline{v}) \quad \text{means} \quad \underline{i} = (g_1(v_1), \ldots, g_{n+p}(v_{n+p}))'.$$

Let the contribution of the current sources $\underline{J}$ to the current through the resistive fundamental cut-sets be denoted by $\hat{\underline{j}}$. We note that

$$\hat{\underline{j}} = \underline{\gamma}_{T(R),\, L(J)}\underline{J} \tag{2}$$

and that $\underline{v}$, the voltages of the resistive branches is given by

$$\underline{v} = \underline{\gamma}'_{T(RE),\, R}\begin{bmatrix} \underline{E} \\ \\ \underline{e} \end{bmatrix}$$

Thus (1) is actually an expression of Kirchoff's current law which states that the sum of the currents through a fundamental cut-set is zero.

Thus, for a given $\underline{E}$ and using (2), (1) can be written as

$$\underline{G}(\underline{e}) + \hat{\underline{j}} = \underline{0} \tag{3}$$

where

$$\underline{G}(\underline{e}) = \underline{\gamma}_{T(R),\, R}\, \mathcal{G}^*\left(\underline{\gamma}_{T(ER),\, R}\begin{bmatrix} \underline{E} \\ \\ \underline{e} \end{bmatrix}\right) \tag{4}$$

is the contribution of the resistive branch currents to the current through the resistive cut-sets.

Hence, to each choice of the vector $\underline{e}$ in the Euclidean space E^n of resistive tree branch voltages there corresponds a vector in the Euclidean space E^n of resistive tree branch cut-set currents, the corresponding being provided by the nonlinear mapping $\underline{G}$.

The solution of the network reduces, therefore, to finding $\underline{e}$ which satisfies (1), given $\underline{E}$ and $\underline{J}$, or by using (3), given $\underline{E}$ and $\hat{\underline{j}}$.

Theorem:

For the network η described above, the solution $\underline{e}$ is given by $\underline{e} = \lim_{m \to \infty} \underline{e}_m$, where

$$\underline{e}_{m+1} = \frac{k_1}{k_2}(-\hat{\underline{j}} - \underline{G}(\underline{e}_m)) + \underline{e}_m \tag{5}$$

and $\underline{e}_o$ is an arbitrary element of E^n. Here k_1 is the smallest conductance in the tree,

$$k_1 = \min_{i=1,\ldots,n} \underline{g}_i \tag{6}$$

where the minimization is carried out on the resistors which are tree branches, and k_2 is given by the square of the mean square matrix norm [4] of the fundamental conductance matrix of a linear network derived from the original network by replacing each nonlinear resistor r_k by a linear resistor whose conductance is equal to $\overline{g}_k$. A bound on this norm is given by

$$k_2 = n^2 \left(\sum_{i=1}^{n+p} \overline{g}_i \right)^2 \tag{7}$$

where n is the number of resistive tree branches and p is the number of links.

Proof:

We appeal to a theorem of I. W. Sandberg [3], a restatement of which in terms of our notation reads as follows: Let $\underline{e}$ and $\overline{\underline{e}}$ be elements of E^n. Denote the inner product of $\underline{e}$ with $\overline{\underline{e}}$ by $(\underline{e}, \overline{\underline{e}})$ and the norm of $\underline{e}$ by $\| \underline{e} \| = (\underline{e}, \underline{e})^{\frac{1}{2}}$.

Let $\underline{G}$ be a mapping from E^n into E^n such that for all $\underline{e}$ and $\overline{\underline{e}}$ in E^n

$$(\underline{G}(\underline{e}) - \underline{G}(\overline{\underline{e}}), \underline{e} - \overline{\underline{e}}) \geq k_1 \| \underline{e} - \overline{\underline{e}} \|^2 \tag{8}$$

$$\| \underline{G}(\underline{e}) - \underline{G}(\overline{\underline{e}}) \|^2 \leq k_2 \| \underline{e} - \overline{\underline{e}} \|^2 \tag{9}$$

where k_1 and k_2 are positive constants. Then for each $\underline{j} \in E^n$ the equation

$$\underline{G}(\underline{e}) = -\underline{j}$$

has a unique solution given by

$$\underline{e} = \lim_{m \to \infty} \{ \frac{k_1}{k_2} (-\underline{j} - \underline{G}(\underline{e}_m)) + \underline{e}_m \},$$

where $\underline{e}_o$ is an arbitrary element of E^n. Hence the proof of our theorem is complete if we can show that (8) and (9) are satisfied for the values of k_1 and k_2 given by (6) and (7). Note also that conditions (8) and (9) imply monotonic convergence, i.e., $\| e - e_m \| < \| e - e_{m-1} \|$, as

$$\begin{aligned}
||\underline{e} - \underline{e}_{m-1}||^2 - ||\underline{e} - \underline{e}_m||^2 &= ||\underline{e} - \underline{e}_{m-1}||^2 - ||\underline{e} - \{k(-\underline{j} - \underline{G}(\underline{e}_{m-1})) + \underline{e}_{m-1}\}||^2 \\
&= ||e - e_{m-1}||^2 - ||\underline{e} - \underline{e}_{m-1} - k(\underline{G}(\underline{e}) - \underline{G}(\underline{e}_{m-1}))||^2 = ||\underline{e} - \underline{e}_{m-1}||^2 \\
&\quad - ||\underline{e} - \underline{e}_{m-1}||^2 + 2k(\underline{e} - \underline{e}_{m-1}, G(\underline{e}) - G(\underline{e}_{m-1})) - k^2 ||G(\underline{e}) - G(e_{m-1})||^2 \\
&= 2k(\underline{e} - \underline{e}_{m-1}, G(\underline{e}) - G(\underline{e}_{m-1})) - k^2 ||\underline{G}(\underline{e}_m) - \underline{G}(\underline{e}_{m-1})||^2
\end{aligned}$$

where $k = \frac{k_1}{k_2}$. But, using (8) and (9),

$$2k(\underline{e}-\underline{e}_{m-1}, \underline{G}(\underline{e})-\underline{G}(\underline{e}_{m-1}))-k^2 \|\underline{G}(\underline{e})-\underline{G}(\underline{e}_{m-1})\| \geq 2kk, \|e-e_{m-1}\|^2$$

$$- k^2 k_2 \|\underline{e}-\underline{e}_{m-1}\|^2 = k^2 k_2 \|\underline{e}-\underline{e}_{m-1}\|^2 \geq 0 ,$$

hence if $\underline{e}_{m-1} \neq \underline{e}$, $\|\underline{e} - \underline{e}_m\| \leq \|\underline{e} - \underline{e}_{m-1}\|$

Let us calculate k_1 first. Let

$$\underline{e} = (e_1, \ldots, e_n)', \qquad \overline{\underline{e}} = (\overline{e}_1, \ldots, \overline{e}_n)'.$$

Then if

$$\hat{\underline{e}} = \begin{bmatrix} \underline{E} \\ \underline{e} \end{bmatrix}, \qquad \hat{\overline{\underline{e}}} = \begin{bmatrix} \underline{E} \\ \overline{\underline{e}} \end{bmatrix},$$

and

$$\underline{v} = \underline{\gamma}'_{T(ER), R}\, \underline{e}, \qquad \overline{\underline{v}} = \underline{\gamma}'_{T(ER), R}\, \overline{\underline{e}},$$

we have that

$$(\underline{G}(\underline{e}) - \underline{G}(\overline{\underline{e}}), \underline{e} - \overline{\underline{e}}) = (\underline{\gamma}_{T(R), R}\, \underline{\mathcal{G}}^*(\underline{v}) - \underline{\gamma}_{T(R), R}\, \underline{\mathcal{G}}^*(\overline{\underline{v}}), \underline{e} - \overline{\underline{e}})$$

$$= (\underline{\gamma}_{T(R), R}(\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\overline{\underline{v}})), \underline{e} - \overline{\underline{e}})$$

$$= (\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\overline{\underline{v}}), \underline{\gamma}'_{T(R), R}(\underline{e} - \overline{\underline{e}}))$$

We shall show that

$$[\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\overline{\underline{v}})]_i\, [\underline{\gamma}'_{T(R),R}(\underline{e} - \overline{\underline{e}})]_i \geq 0$$

for all i, where i denotes the i-th component of the vector.

We note that since each element of the tree is an element of its own fundamental cut-set and no others,

$$\underline{\gamma}_{T(ER), R} = \left[\begin{array}{c:c} \underline{0} & \underline{\gamma}_{T(E), L(R)} \\ \hdashline \multicolumn{2}{c}{\underline{\gamma}_{T(R), R}} \end{array}\right]$$

where $\underline{0}$ is the m x n null matrix, and

$$\underline{\gamma}_{T(R), R} = (\underline{I} \;\vdots\; \underline{\gamma}_{T(R), L(R)}), \tag{10}$$

where $\underline{I}$ is the unit matrix.

Thus

$$\underline{v} = \underline{\gamma}'_{T(ER), R} \begin{pmatrix} \underline{E} \\ \underline{e} \end{pmatrix} = \underline{\gamma}'_{T(R), R}\, \underline{e} + \gamma'_{T(E), R}\, \underline{E} ,$$

and

$$\underline{v} - \bar{\underline{v}} = \Upsilon'_{T(R),\,R}(\underline{e} - \bar{\underline{e}}).$$

Since $\underline{\mathcal{G}}^*(\underline{v}) = \begin{bmatrix} g_1 & & \\ & \ddots & \\ & & g_{n+p} \end{bmatrix}(\underline{v})$, we have

$$[\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\bar{\underline{v}})]_i\ [\Upsilon'_{T(R),\,R}(\underline{e} - \bar{\underline{e}})]_i = [\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\bar{\underline{v}})]_i\ [\underline{v} - \bar{\underline{v}}]_i$$

$$\geq \underline{g}_i\,(v_i - \bar{v}_i)^2 \geq 0$$

in consequence of conditions (a) and (b) imposed on the resistors.

From (10) it follows that for $i \leq n$, namely resistors which are tree branches $v_i = e_i$ and $\bar{v}_i = \bar{e}_i$.

Thus,

$$(\underline{G}(\underline{e}) - \underline{G}(\bar{\underline{e}}),\ \underline{e} - \bar{\underline{e}}) = \left(\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\bar{\underline{v}}),\ \Upsilon'_{T(R),\,R}(\underline{e} - \bar{\underline{e}})\right)$$

$$\geq \sum_{i=1}^{n} [\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\bar{\underline{v}})]_i\ {}^*[\Upsilon'_{T(R),\,R}(\underline{e} - \bar{\underline{e}})]_i$$

$$= \sum_{i=1}^{n} [g_i(e_i) - g_i(\bar{e}_i)] \cdot [e_i - \bar{e}_i]$$

$$\geq \sum_{i=1}^{n} \underline{g}_i\,(e_i - \bar{e}_i)^2$$

$$\geq \Big\{\sum_{i=1}^{n} (e_i - \bar{e}_i)^2\Big\} \min_{j=1,\ldots,n} \underline{g}_j$$

$$= k_1 \|\underline{e} - \bar{\underline{e}}\|^2\,,$$

which proves (8).

To demonstrate (9), we use relation (4) and find that

$$\|\underline{G}(\underline{e}) - \underline{G}(\bar{\underline{e}})\|^2 = \|\Upsilon_{T(R),\,R}\,\underline{\mathcal{G}}^*(\underline{v}) - \Upsilon_{T(R),\,R}\,\underline{\mathcal{G}}^*(\bar{\underline{v}})\|^2$$

$$= \|\Upsilon_{T(R),\,R}\,(\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\bar{\underline{v}}))\|^2.$$

For $i = 1, \ldots, n+p$, define $\hat{g}_i$ to be

$$\hat{g}_i = \frac{g_i(v_i) - g_i(\bar{v}_i)}{v_i - \bar{v}_i} \qquad \text{if } v_i \neq \bar{v}_i$$

$\hat{g}_i =$ any number between $\underline{g}_i$ and $\bar{g}_i$ $\qquad$ if $v_i = \bar{v}_i$

then $\underline{g}_i \le \hat{g}_i \le \overline{g}_i$, and

$$\underline{\mathcal{Y}}_{T(R),R}(\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\overline{\underline{v}})) = \underline{\mathcal{Y}}_{T(R),R}\begin{bmatrix} g_1(v_1) - g_1(\overline{v}_1) \\ \vdots \\ g_{n+p}(v_{n+p}) - g_1(\overline{v}_{n+p}) \end{bmatrix}$$

$$= \underline{\mathcal{Y}}_{T(R),R}\begin{bmatrix} (v_1 - \overline{v}_1)\hat{g}_1 \\ \vdots \\ (v_{n+p} - \overline{v}_{n+p})\hat{g}_{n+p} \end{bmatrix}$$

We note that the $\hat{g}_1$'s are functions and strictly speaking the notation should have been $\hat{g}_i(v_i, \overline{v}_i)$. The abbreviated notation $\hat{g}_i$ is used for simplicity since the value of v_i and $\overline{v}_i$ is clear from the content of the equations.

Let

$$\hat{\underline{G}} = \begin{bmatrix} \hat{g}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{g}_{n+p} \end{bmatrix}, \quad \overline{\underline{G}} = \begin{bmatrix} \overline{g}_1 & & 0 \\ & \ddots & \\ 0 & & \overline{g}_{n+p} \end{bmatrix}.$$

Then

$$\begin{aligned} \underline{\mathcal{Y}}_{T(R),R}(\underline{\mathcal{G}}^*(\underline{v}) - \underline{\mathcal{G}}^*(\overline{\underline{v}})) &= \underline{\mathcal{Y}}_{T(R),R}\hat{\underline{G}}(\underline{v} - \overline{\underline{v}}) \\ &= \underline{\mathcal{Y}}_{T(R),R}\hat{\underline{G}}\left(\underline{\mathcal{Y}}'_{T(ER),R}\hat{\underline{e}} - \underline{\mathcal{Y}}'_{T(ER),R}\hat{\overline{\underline{e}}}\right) \\ &= \underline{\mathcal{Y}}_{T(R),R}\hat{\underline{G}}\underline{\mathcal{Y}}'_{T(ER),R}(\hat{\underline{e}} - \hat{\overline{\underline{e}}}) \\ &= \underline{\mathcal{Y}}_{T(R),R}\hat{\underline{G}}\underline{\mathcal{Y}}'_{T(R),R}(\underline{e} - \overline{\underline{e}}) \end{aligned}$$

The last line is implied by

$$\hat{\underline{e}} = \begin{pmatrix} \underline{E} \\ \underline{e} \end{pmatrix} \quad \text{and} \quad \hat{\underline{e}} - \hat{\overline{\underline{e}}} = \begin{pmatrix} \underline{0} \\ \underline{e} - \overline{\underline{e}} \end{pmatrix}.$$

Hence

$$\begin{aligned} \| \underline{G}(\underline{e}) - \underline{G}(\overline{\underline{e}}) \|^2 &= \| \underline{\mathcal{Y}}_{T(R),R}\hat{\underline{G}}\underline{\mathcal{Y}}'_{T(R),R}(\underline{e} - \overline{\underline{e}}) \|^2 \\ &\le \| \underline{\mathcal{Y}}_{T(R),R}\hat{\underline{G}}\underline{\mathcal{Y}}'_{T(R),R} \|^2 \, \| \underline{e} - \overline{\underline{e}} \|^2 \qquad (11) \end{aligned}$$

It remains to be shown that

$$\| \underline{\mathcal{Y}}_{T(R),R}\hat{\underline{G}}\underline{\mathcal{Y}}'_{T(R),R} \|^2 \le \| \underline{\mathcal{Y}}_{T(R),R}\overline{\underline{G}}\underline{\mathcal{Y}}'_{T(R),R} \|^2 \qquad (12)$$

Denote the matrix on the l.h.s. of (12) by $\underline{A}$ and the matrix on the r.h.s. by $\underline{B}$. Note that for any given $\underline{e}$ and $\underline{e}$, the l.h.s. is the norm of a conductance matrix of a linear network which was derived from the original network by replacing each resistor r_i by a linear resistor whose conductance is $\hat{g}_i$. Similarly the r.h.s. is a conductance matrix of a network obtained by replacing each nonlinear resistor r_i with a linear resistor whose conductance is $\bar{g}_i$. Being conductance matrices, both $\underline{A}$ and $\underline{B}$ are symmetric and positive definite. Since $g_i \leq \bar{g}_i$ we can consider $\underline{C} = \underline{B} - \underline{A}$ as another conductance matrix with conductance $\bar{g}_i - \hat{g}_i$. Therefore $\underline{C}$ is also symmetric and nonnegative definite. Since $\underline{A}$ is symmetric, $\|\underline{A}\| = \lambda_A$ where λ_A is the largest eigenvalue of A. Let the row vector $\underline{x}_1$ be the normalized eigenvector of $\underline{A}$ which corresponds to λ_A. Then

$$\underline{x}_1 \underline{Ax}'_1 + \underline{x}_1 \underline{Cx}'_1 = \underline{x}_1 \underline{Bx}'_1$$

$$\lambda_A + \underline{x}_1 \underline{Cx}'_1 \leq \sup_{\|x_1\|=1} \underline{x}_1 \underline{Bx}'_1 = \lambda_B, \tag{13}$$

where $\underline{x}$ denotes a row vector of appropriate dimensions. As $\underline{C}$ is nonnegative definite and both λ_A and λ_B are real and positive it follows that

$$\lambda_A \leq \lambda_B$$

which concludes the proof of the theorem. To prove (7) consider the following: A bound on a s x s matrix is given by

$$\|\underline{A}\| \leq \left[\sum_{i=1}^{s} \sum_{j=1}^{s} |a_{ij}|^2 \right]^{1/2} \tag{14}$$

Hence, using (12) through (14),

$$\|\underline{\gamma}_{T(R),\,R}\, \underline{G}\, \underline{\gamma}_{T(R),\,R}\|^2 \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sum_{k=1}^{n+p} \bar{g}_k \right)^2 = n^2 \left(\sum_{k=1}^{n+p} \bar{g}_k \right)^2 .$$

Thus,

$$\|\underline{G}(\underline{e}) - \underline{G}(\bar{\underline{e}})\|^2 \leq k_2 \|\underline{e} - \bar{\underline{e}}\|^2$$

where

$$k_2 = n^2 \left(\sum_{k=1}^{n+p} \bar{g}_k \right)^2 ,$$

proving (7), and hence completing the proof of the theorem.

In practical applications the calculation of k_2 is performed using relations (12) and (14). In view of these relations, we could replace k_2 in the theorem by

$$k_2 = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij}^2, \qquad (15)$$

where

$$B = (b_{ij}) = \gamma_{T(R),\, L(R)} \underline{\bar{G}}\, \underline{\gamma}'_{T(R),\, L(R)_{ij}}$$

and obtain a tighter bound on k_2 than (7) indicates.

It should be noted that the norm of the matrix on the r.h.s. of (12) can be bounded by using formulas other than (14) with somewhat different results for k_2.

Computer Program

Two computer programs were written in FORTRAN to solve circuits of the type considered by means of the iteration method discussed above. The first program is a direct application of relation (4), while the second, which we call the modified program, incorporates a heuristic modification for reduction of solution time.

The input to these programs consists of the fundamental cut-set matrix and the resistor characteristics which are given as piecewise linear functions.[*] The output is the set of branch voltages.

In the cases tested by the authors the heuristic refinement was found to reduce the time required for solution by 25 Percent to 95 Percent. The refinement is based on the idea that although (5) specified a "jump" that always reduces the norm of the error, often a larger "jump" can be taken. Thus at the n-th step the program tries a larger "jump", compares the resulting error with the one resulting from (5) and chooses the point which corresponds to a smaller error to be $\underline{e}_{n+1}$. The following is a detailed description of the heuristic modification.

Let $\underline{e}_i$ be the vector of resistive tree branch voltages specified after i iterations, and let $\delta\underline{e}_i$ be the vector of voltage changes calculated by the simple iteration method, i.e.,

$$\delta\underline{e}_i = c(-\hat{\underline{j}} - \underline{G}(\underline{e}_i))$$

where c is the iteration constant $c = k_1/k_2$ of (5) (see Fig. 1).

---

[*] We note that while piecewise representations of the resistor characteristics are used here, the iterative method is limited only by requirements (a), (b) stated at the beginning of the paper.

Consider that point $\underline{j}^*$ on the line joining $\underline{G}(\underline{e}_i)$ and $\underline{G}(\underline{e}_i + \delta\underline{e}_i)$ which is closest to $-\hat{\underline{j}}$, and perform a one-step iteration with an iteration constant c^* given by

$$c^* = c \, \frac{\|\underline{j}^* - \underline{G}(\underline{e}_i)\|}{\|\underline{G}(\underline{e}_i+\delta\underline{e}_i) - \underline{G}(\underline{e}_i)\|}\,.$$

Denote the result of this step by

$$\Delta\underline{e}_i = c^*(-\hat{\underline{j}} - G(\underline{e}_i)),$$

The value of $\underline{e}_{i+1}$ is chosen as follows: Compare $L_1 = \|-\hat{\underline{j}} - \underline{G}(\underline{e}_i+\delta\underline{e}_i)\|$ with $L_2 = \|-\hat{\underline{j}} - \underline{G}(\underline{e}_i+\Delta\underline{e}_i)\|$. Then set

$$\underline{e}_{i+1} = \begin{cases} \underline{e}_i \quad \delta\underline{e}_i, & \text{if } L_1 \leq L_2 \\ \underline{e}_i + \Delta\underline{e}_i, & \text{if } L_1 > L_2 \end{cases}$$

In effect, this means that one iteration of the modified program incorporates two iterations of the type performed in the regular program with a selection of the better result.

For all circuits for which solutions were obtained using both programs, it was found that the heuristic modification significantly reduced the time required for solution. However, it should be noted that for any particular iteration L_1 need not exceed L_2, i.e., the one-step iteration using c^* need not always be better than that using c.

Summary and Conclusions

A direct iteration method for solving networks composed of independent sources and nonlinear resistors has been studied in detail and a computation algorithm devised for realizing the method. Convergence of the algorithm for all initial conditions was proved. Two methods of computing the iteration coefficient were presented and computer programs utilizing these methods written. The most satisfactory convergence rate was obtained when the described heuristic combination of the two methods was employed.

Acknowledgement

The authors wish to thank Dr. I. W. Sandberg for helpful discussions and Drs. A. J. Goldstein and J. F. Kaiser for reading and commenting on the manuscript.

BIBLIOGRAPHY

1. Duffin, R. J., Nonlinear Networks II, Bulletin of the American Mathematical Society, 53, pp. 963-971, October 1947.

2. Desoer, C. A. and Katzenelson, J., Nonlinear RLC Networks, Bell System Technical Journal, 44, pp. 161-198, January 1965.

3. Sandberg, I. W., On Properties of Some Systems that Distort Signals - I, Bell System Technical Journal, 42, pp. 2033-2046, September 1963.

4. Faddeeva, V. N., Computational Methods of Linear Algebra, New York, Dover, 1959, p. 60.

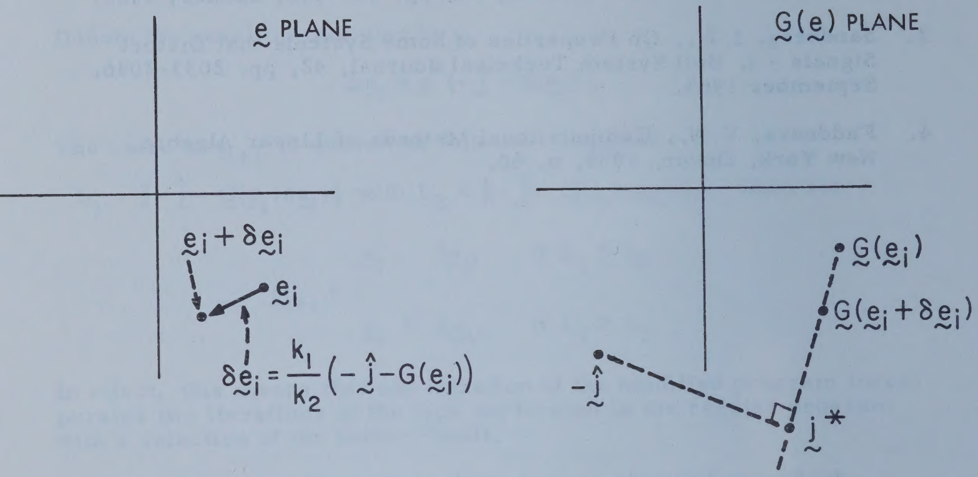

$\underset{\sim}{e}$ PLANE
$\underset{\sim}{G}(\underset{\sim}{e})$ PLANE
$\underset{\sim}{e}_i + \delta \underset{\sim}{e}_i$
$\underset{\sim}{e}_i$
$\delta \underset{\sim}{e}_i = \frac{k_1}{k_2}(-\hat{\underset{\sim}{j}} - G(\underset{\sim}{e}_i))$
$-\hat{\underset{\sim}{j}}$
$\underset{\sim}{G}(\underset{\sim}{e}_i)$
$\underset{\sim}{G}(\underset{\sim}{e}_i + \delta \underset{\sim}{e}_i)$
$\underset{\sim}{j}^*$

Fig. 1 A Heuristic Modification for Improvement of Convergence Rate.

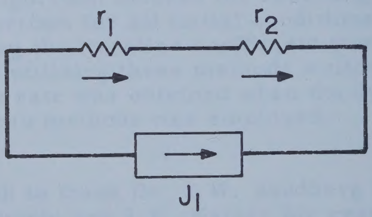


Fig. 2 A Simple Example.

COMPREHENSIVE ACTIVE NETWORK ANALYSIS BY DIGITAL COMPUTER–A STATE-SPACE APPROACH

CHRISTOPHER POTTLE
School of Electrical Engineering, Cornell University, Ithaca, N.Y.

ABSTRACT

An extension of Bashkow's A-matrix to the active case provides a convenient description of the most general type of linear network. A recursive scheme developed by Dervisoglu for obtaining this matrix yields an algorithm ideally suited for implementation on a digital computer. This paper presents a modification of his scheme, forming the basis of a comprehensive computer analysis program for general active networks. The advantages of this approach include the capability of obtaining several network functions simultaneously, of working in either the time or frequency domain, and of adjusting parameters without recalculation of the entire network.

A. INTRODUCTION

Countless techniques now exist for carrying out the analysis of linear networks in an orderly fashion. Many of these have their own peculiar advantages, and many lend themselves readily to implementation on a digital computer. Needless to say, any new scheme must offer some considerable advantages to make its implementation worthwhile. Present network analysis techniques may be broadly classified into the three groups which follow.

1. Loop or Nodal Analysis Methods

These classical techniques work very well in those cases where the analysis desired is the frequency response of a passive network without many parameter changes. The true order of an active network may be difficult to determine by these methods, and the inversion of a matrix of rational functions to obtain transfer functions explicitly is an unpleasant task at best. Complete recalculation of the network is required when a parameter is altered.

2. Topological Formulae

These methods generally require the generation and listing of the trees of the network graph. They offer the possibility of obtaining transfer functions in terms of literal element names instead of values, and therefore the effect of changing many element values is almost trivially obtained. It would seem that these methods are restricted to this sort of application, since the computer time involved for the initial analysis of the network is far higher than for other schemes.

3. The State-Space Approach

Rather more recently, several techniques have appeared to describe a network in terms of a set of independent first-order differential equations, the unknowns (state variables) being a subset of the capacitor voltages and inductor currents in the network.[1-4] In this

---

This research was supported by the Ford Foundation through a grant to the College of Engineering, Cornell University.

formulation, the order of the network is immediately apparent, being the number of equations required to describe the network. For passive networks, this number is easily determined by inspection,[5] but the presence of active elements may serve to raise or lower this number. Indeed, most methods of active network synthesis use active devices in a way that lowers the over-all order of the network. Obtaining the network equations in this form has several advantages:

(a) Transfer functions are easily found using an algorithm that makes particular use of the form of the equations to avoid the usual problem of rational function matrix inversion.[6,7]
(b) Time responses can be easily found by Taylor-series approximations to the transition matrix instead of inverse-transform techniques requiring factoring of the characteristic polynomial.[8]
(c) The entire process consists of simple matrix manipulations.
(d) These methods seem to be the most amenable to extension to nonlinear and time-varying networks.

The digital computer algorithm described in this paper belongs to this third group. It can obtain the state equations of any active linear network, and its scheme is based largely on the work of Dervisoglu, who seems to have been the first to derive a general solution for the active case.[9] The active networks used consist of the usual passive elements of resistance, inductance, mutual inductance, capacitance, and mutual capacitance, with the addition of the dependent voltage (or current) source controlled by a branch voltage (or current). Multiterminal devices, which are often used in active network analysis, are easily represented as a collection of the elements mentioned above. One sample network is used throughout to illustrate each point as it is presented. This circuit, shown in Fig. 1, uses a negative impedance converter to obtain a non-positive—real driving-point admittance—in this case a quadratic polynomial.[10]

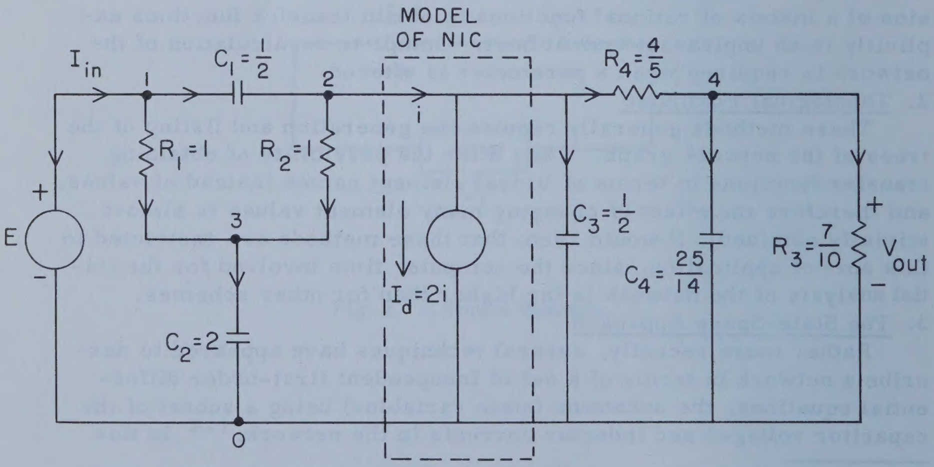


Fig. 1. Active Network Used in Example.

B. THE PROPER TREE AND FUNDAMENTAL LOOP MATRIX

As in other procedures, this approach first requires the formation of the fundamental loop matrix (or its negative transpose, the fundamental cut-set matrix) with respect to a "proper" tree;[1] that is, a tree that contains all voltage sources as tree branches, all current sources as links, and as many capacitive tree branches and inductive links as possible. This tree may be found and the loop matrix determined by an easily mechanized procedure. The incidence matrix of the network is first constructed. Its n-1 rows refer to nodes, and its columns to branches. Its columns should be arranged in the order: voltage sources, capacitors, resistors, inductors, and current sources. Each row expresses the Current Law at one node, which implies the matrix equation:

$$A\,\underline{i} = \underline{0} \quad . \tag{1}$$

In our example, the incidence matrix is:

$$\begin{array}{c|cccccccccc}
\text{Node} & E & C_1 & C_2 & C_3 & C_4 & R_1 & R_2 & R_3 & R_4 & I_d \\
\hline
1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
2 & 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\
3 & 0 & 0 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 & 0
\end{array} \quad .$$

The next step consists of performing row operations on the incidence matrix in an attempt to reduce (1) to the form:

$$[\,-I \mid F\,]\,\underline{i} = \underline{0} \quad . \tag{2}$$

This equation states that, if all currents identified with columns of F (link currents) except one are set to zero, then that current flows in certain tree branches (columns of -I); i.e., F is the fundamental loop matrix of the circuit with respect to the tree defined by the first n-1 columns. When this reduction is attempted, it may not be possible to obtain the desired result because the first n-1 branches chosen as columns do not form a tree. In any case, the matrix should be reduced to row echelon form. In our example, this reduction results in the matrix:

$$\begin{array}{cccccccccc}
E & C_1 & C_2 & C_3 & C_4 & R_1 & R_2 & R_3 & R_4 & I_d \\
\hline
-1 & 0 & 0 & -1 & 0 & -1 & -1 & 0 & -1 & -1 \\
0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & -1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 1 & 0
\end{array} \quad .$$

The proper tree consists of E, C_1, C_2, and C_4, as shown in Fig. 2.

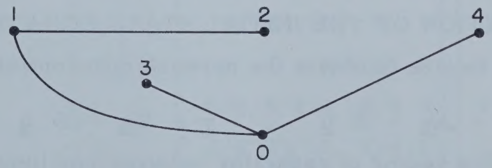


Fig. 2. Proper Tree for Example.

The row-reduced matrix may now be brought into the desired form by a transposition of columns. This transposition should be done in such a way that the branches are ordered in the following manner:

Current symbol	Branch	Voltage symbol
i_d	Dependent Voltage Sources	V_d
i_e	Independent Voltage Sources	E
i_c	Tree Capacitors	V_c
i_l	Tree (excess) Inductors	V_l
i_r	Tree Resistors	V_r
I_r	Link Resistors	v_r
I_c	Link (excess) Capacitors	v_c
I_l	Link Inductors	v_l
J	Independent Current Sources	v_j
I_d	Dependent Current Sources	v_d

After this rearrangement, the relations between tree branch currents and link currents involving F and between link voltages and tree branch voltages involving $-F^t$ become:

$$\begin{bmatrix} i_d \\ i_e \\ i_c \\ i_l \\ i_r \end{bmatrix} = \begin{bmatrix} F_{11} & F_{12} & F_{13} & F_{14} & F_{15} \\ F_{21} & F_{22} & F_{23} & F_{24} & F_{25} \\ F_{31} & F_{32} & F_{33} & F_{34} & F_{35} \\ 0 & 0 & F_{43} & F_{44} & F_{45} \\ F_{51} & 0 & F_{53} & F_{54} & F_{55} \end{bmatrix} \begin{bmatrix} I_r \\ I_c \\ I_l \\ J \\ I_d \end{bmatrix} ; \quad \begin{bmatrix} v_r \\ v_c \\ v_l \\ v_j \\ v_d \end{bmatrix} = [-F^t] \begin{bmatrix} V_d \\ E \\ V_c \\ V_l \\ V_r \end{bmatrix} . \qquad (3)$$

The zero submatrices are required by the way excess elements are defined. For the example F is:

$$\begin{array}{c|cccccc|} & R_1 & R_2 & R_3 & R_4 & C_3 & I_d \\ \hline i_e & -1 & -1 & 0 & -1 & -1 & -1 \\ i_{c1} & 0 & 1 & 0 & 1 & 1 & 1 \\ i_{c2} & 1 & 1 & 0 & 0 & 0 & 0 \\ i_{c4} & 0 & 0 & -1 & 1 & 0 & 0 \end{array} .$$

These matrix operations are easily and speedily carried out on a binary computer by logical operations on packed data, since at no time is any result obtained other than 1, 0, or -1.[11] If zero rows turn up at the bottom of F during reduction, the graph of the network is not connected. If any voltage sources turn up as links, or current sources as trees, the network contains a loop of voltage sources or cut-set of current sources and thus has no solution.

C. FORMATION OF THE INITIAL STATE EQUATIONS

Eventually we desire to obtain the network differential equations in the form:

$$\dot{\underline{x}} = A\underline{x} + B\,\underline{u} \quad , \quad \underline{y} = C\underline{x} + D\,\underline{u} \quad , \qquad (4)$$

where $\underline{x}$ is the state vector of capacitor voltages and inductor currents, $\underline{u}$ is the vector of independent sources, and $\underline{y}$ is the vector of desired output quantities. The first step in this process is to express the dependent sources and output quantities in terms of tree branch voltages and link currents. Wherever expressions involving lower case variables are encountered, these variables are removed by consulting F or $-F^t$ in (3). The result has the form:

$$\begin{array}{l} [P]\begin{bmatrix} V_d \\ I_d \end{bmatrix} = \begin{bmatrix} \quad Q \quad \end{bmatrix} , \qquad (5) \\ \\ \underline{y} = \left[C_1 \,\vdots\, C_2 \,\vdots\, C_3 \right] \begin{bmatrix} E \\ V_c \\ V_l \\ V_r \\ I_r \\ I_c \\ I_l \\ J \end{bmatrix} + \begin{bmatrix} C_4 \end{bmatrix} \begin{bmatrix} V_d \\ I_d \end{bmatrix} . \qquad (6) \end{array}$$

The matrix P may differ from the identity matrix because of the dependence of one controlled source upon another. The solution of (5) is now obtained and substituted into (6). For the example we obtain:

$$I_d = 2i = 2\left(i_{c1} - I_{r2}\right) = 2\left(I_{r2} + I_{r4} + I_{c3} + I_d - I_{r2}\right) = -2\,I_{r4} - 2I_{c3} \, , \qquad (7)$$

$$I_{in} = -i_e = I_{r1} + I_{r2} + I_{r4} + I_{c3} + I_d = I_{r1} + I_{r2} - I_{r4} - I_{c3} \, , \qquad (8)$$

$$V_{out} = V_{c_4} \, . \qquad (9)$$

The possibility of a singular P matrix is discussed in the next section.

Similarly, the resistive variables may be expressed in terms of the remaining upper case quantities by introducing the resistive element values. Again the possibility of a singular matrix arises, but only in active networks (see the next section). This result may now be used to remove the dependence on resistive variables in (5) - (9). We have now obtained expressions of the form:

$$\begin{array}{c} \begin{bmatrix} V_d \\ I_d \end{bmatrix} \\ \begin{bmatrix} V_r \\ I_r \end{bmatrix} \\ \begin{bmatrix} y \end{bmatrix} \end{array} \begin{array}{c} = \\ \\ = \\ \\ = \end{array} \begin{array}{c} \begin{bmatrix} \qquad Q \qquad \end{bmatrix} \\ \begin{bmatrix} \qquad T \qquad \end{bmatrix} \\ \left[D_1 \,\vdots\, C_1 \,\vdots\, C_2 \,\vdots\, C_3 \,\vdots\, C_4 \,\vdots\, D_2 \right] \end{array} \begin{bmatrix} E \\ V_c \\ V_l \\ I_c \\ I_l \\ J \end{bmatrix} \begin{array}{c} , \\ \\ , \\ \\ . \end{array} \qquad (10)$$

The result for our example is:

$$\begin{array}{c} \begin{bmatrix} I_d \end{bmatrix} \\ \begin{bmatrix} I_{r_1} \\ I_{r_2} \\ I_{r_3} \\ I_{r_4} \end{bmatrix} \\ \begin{bmatrix} I_{in} \\ V_{out} \end{bmatrix} \end{array} = \begin{array}{c} \begin{bmatrix} -5/2 & 5/2 & 0 & 5/2 & -2 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 10/7 & 0 \\ 5/4 & -5/4 & 0 & -5/4 & 0 \end{bmatrix} \\ \begin{bmatrix} 3/4 & 1/4 & -2 & 5/4 & -1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{bmatrix} E \\ V_{c_1} \\ V_{c_2} \\ V_{c_4} \\ I_{c_3} \end{bmatrix} . \qquad (11)$$

A set of equations relating the currents and voltages of the reactive elements can now be written by consulting the F and $-F^t$ matrices. These equations eventually become the initial state equations of the network. Thus

$$\begin{bmatrix} i_c\\ i_l\\ v_c\\ v_l \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & & F_{32} & F_{33}\\ & & 0 & F_{43}\\ \hline -F_{32}^t & 0 & 0 & \\ -F_{33}^t & -F_{43}^t & & \end{array}\right] \begin{bmatrix} V_c\\ V_l\\ I_c\\ I_l \end{bmatrix} + \left[\begin{array}{c|c} 0 & F_{34}\\ & F_{44}\\ \hline -F_{22}^t & 0\\ -F_{23}^t & \end{array}\right] \begin{bmatrix} E\\ J \end{bmatrix} + \left[\begin{array}{c|c} 0 & F_{31}\\ & 0\\ \hline 0 & \\ -F_{53}^t & 0 \end{array}\right] \begin{bmatrix} V_r\\ I_r \end{bmatrix} + \left[\begin{array}{c|c} 0 & F_{35}\\ & F_{45}\\ \hline -F_{12}^t & 0\\ -F_{13}^t & \end{array}\right] \begin{bmatrix} V_d\\ I_d \end{bmatrix} . \qquad (12)$$

Using (10), the resistive quantities and controlled sources are removed and the rows rearranged to obtain:

$$\begin{bmatrix} i_c\\ v_l\\ i_l\\ v_c \end{bmatrix} = \left[\begin{array}{c|c|c|c} A_1 & A_2 & A_3 & A_4 \end{array}\right] \begin{bmatrix} V_c\\ V_l\\ I_c\\ I_l \end{bmatrix} + \begin{bmatrix} B \end{bmatrix} \begin{bmatrix} E\\ J \end{bmatrix} . \qquad (13)$$

These matrices become for our example:

$$A = \begin{bmatrix} 1/4 & -1 & 5/4 & -1\\ -1 & -2 & 0 & 0\\ -5/4 & 0 & -75/28 & 0\\ -1 & 0 & 0 & 0 \end{bmatrix}; \quad B = \begin{bmatrix} -1/4\\ 2\\ 5/4\\ 1 \end{bmatrix} . \qquad (14)$$

At this point the branch equations for the reactive elements themselves are introduced:

$$\begin{bmatrix} V_l\\ v_l \end{bmatrix} = \left[\begin{array}{c} L_1\\ \hline L_2 \end{array}\right] \begin{bmatrix} \dot{i}_u\\ \dot{I}_l \end{bmatrix} ; \begin{bmatrix} i_c\\ I_c \end{bmatrix} = \left[\begin{array}{c} H_1\\ \hline H_2 \end{array}\right] \begin{bmatrix} \dot{V}_c\\ \dot{v}_c \end{bmatrix} . \qquad (15)$$

Some reshuffling is still in order, since the state variables for the excess elements are of the wrong sort (capacitor currents and inductor voltages). Writing these quantities in terms of the derivatives of the proper variables using (15) and expanding the vector of the left-hand side of (15), we obtain the initial state equations:

$$\underset{S}{\left[\begin{array}{c|c} H_1 & \\ \hline & L_2\\ \cline{2-2} -A_3H_2 & -A_2L_1\\ & \end{array}\right]} \underset{\dot{\underline{x}}}{\begin{bmatrix} \dot{V}_c\\ \dot{v}_c\\ \dot{i}_l\\ \dot{I}_l \end{bmatrix}} = \underset{A}{\left[\begin{array}{c|c|c|c} & 0 & 0 & \\ A_1 & 0 & 0 & A_4\\ \cline{3-3} & 0 & -I & \\ \cline{2-3} & -I & 0 & \end{array}\right]} \underset{\underline{x}}{\begin{bmatrix} V_c\\ v_c\\ i_l\\ I_l \end{bmatrix}} + \underset{B}{\begin{bmatrix} \\ B\\ \\ \end{bmatrix}} \underset{\underline{u}}{\begin{bmatrix} E\\ J \end{bmatrix}} , \qquad (16)$$

It is easily shown that, in the passive case, a number of rows at the bottom of the S matrix equal to the number of excess elements are zero.

The expression for the output variables in (10) also contains the incorrect type of state variables for excess elements. A modification yields:

$$\begin{bmatrix} y \end{bmatrix} = \left[\begin{array}{c|c} C_3H_2 & C_2L_1 \end{array}\right] \dot{\underline{x}} + \left[\begin{array}{c|c|c|c} C_1 & 0 & 0 & C_4 \end{array}\right] \underline{x} + \left[\begin{array}{c|c} D_1 & D_2 \end{array}\right] \underline{u} ,$$

$$\underline{y} = CT \; \dot{\underline{x}} + C \; \underline{x} + D \; \underline{u} . \qquad (17)$$

For the purposes of the reduction described in the next section, the matrices involved in (16) and (17) are most easily regarded as submatrices of one large partitioned matrix:

$$\left[\begin{array}{c|c|c} CT & C & D \\ \hline S & A & B \end{array}\right] . \tag{18}$$

This matrix is, for our example:

$$\left[\begin{array}{cccc|cccc|c} 0 & 0 & 0 & -1/2 & 1/4 & -2 & 5/4 & 0 & 3/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1/2 & 0 & 0 & 1/2 & 1/4 & -1 & 5/4 & 0 & -1/4 \\ 0 & 2 & 0 & 0 & -1 & -2 & 0 & 0 & 2 \\ 0 & 0 & 25/14 & 0 & -5/4 & 0 & -75/28 & 0 & 5/4 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 1 \end{array}\right] . \tag{19}$$

D. REDUCTION OF STATE EQUATIONS

The state equation (16) would be in final form if S were an identity matrix. Since this is not generally so, we proceed by reducing the partitioned matrix $[S \,\vdots\, A \,\vdots\, B]$ to row echelon form.[9] Several possibilities that now arise are discussed in turn:

1. S becomes an identity matrix. In this case, the state equations are in the desired form and the reduction is complete.

2. S and A contain at least one zero row. B should have at least one nonzero element in this row; otherwise, the presence of a completely superfluous state equation is indicated. But such a row indicates that a relation exists between independent driving sources. In this case, the network has no solution. The physical interpretation of this event is that a dependent source or sources inside the network are causing a voltage source to see a short circuit, or a current source an open circuit.

3. S but not A contains at least one zero row. In this case, a set of the state variables are linearly dependent, and at least one state variable can be removed. In the passive case, the state variables of all excess elements can be removed, together with one state variable for every unity-coupled transformer. The k^{th} state variable is removed by row operations to reduce all elements in the k^{th} column of A to zero. Since the C matrix lies directly above A, the removal of the variable from the output equations is easily accomplished at the same time. The derivative of this state variable must also be removed from the left-hand side of the equation, if it appears. This removal is accomplished by differentiating the expression relating this variable to the others and again using row operations to reduce the k^{th} column of S and CT to zero. Notice that in this way derivatives of the source quantities may appear. When the process is complete, the m rows that were zero in S may be removed along with the m columns of CT, S, C, and A corresponding to the m dependent state variables removed. The process of reduction to row-echelon form is now repeated until a diagonal S matrix is obtained. At this point, the relation $A\,\underline{x} + B\,\underline{u}$ is substituted for $\dot{\underline{x}}$ in the output equation (17), and the final equations in the form of (4) are obtained. Note, however, that derivatives of the inputs may be present. In fact, if m is the number of reactive elements, n the number of final state variables, and k the number of input derivatives present:

$$k \leq m - n \quad . \tag{20}$$

For passive networks, it may be shown that

$$k = \min(m - n,\ 1) \quad . \tag{21}$$

In our example, the matrix $[S \,\vdots\, A \,\vdots\, B]$ in expression (19) is already in reduced form and indicates the linear dependence:

$$v_{c_3} = E - V_{c_1} \quad ; \tag{22}$$

therefore

$$\dot{v}_{c_3} = \dot{E} - \dot{V}_{c_1} \quad . \tag{23}$$

The quantities v_{c_3} and $\dot{v}_{c_3}$ are now removed to yield:

$$\begin{aligned} [y] &= \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{V}_{c_1} \\ \dot{V}_{c_2} \\ \dot{V}_{c_4} \end{bmatrix} + \begin{bmatrix} 1/4 & -2 & 5/4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_{c_1} \\ V_{c_2} \\ V_{c_4} \end{bmatrix} + \begin{bmatrix} 3/4 & -1/2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} E \\ \dot{E} \end{bmatrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{V}_{c_1} \\ \dot{V}_{c_2} \\ \dot{V}_{c_4} \end{bmatrix} &= \begin{bmatrix} 1/2 & -2 & 5/2 \\ -1/2 & -1 & 0 \\ -7/10 & 0 & -3/2 \end{bmatrix} \begin{bmatrix} V_{c_1} \\ V_{c_2} \\ V_{c_4} \end{bmatrix} + \begin{bmatrix} -1/2 & -1 \\ 1 & 0 \\ 7/10 & 0 \end{bmatrix} \begin{bmatrix} E \\ \dot{E} \end{bmatrix} \quad . \end{aligned} \tag{24}$$

This equation is again in reduced form and indicates a linear dependence:

$$1/2\ V_{c_1} - 2\ V_{c_2} + 5/2\ V_{c_4} - 1/2\ E - \dot{E} = 0 \quad . \tag{25}$$

In this case, any one of the state variables may be removed. If V_{c_4} is chosen and removed, the result is:

$$\begin{aligned} \underline{y} &= \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{V}_{c_1} \\ \dot{V}_{c_2} \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ -1/5 & 4/5 \end{bmatrix} \begin{bmatrix} V_{c_1} \\ V_{c_2} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1/5 & 2/5 & 0 \end{bmatrix} \begin{bmatrix} E \\ \dot{E} \\ \ddot{E} \end{bmatrix} \\ \begin{bmatrix} 0 & 1 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} \dot{V}_{c_1} \\ \dot{V}_{c_2} \end{bmatrix} &= \begin{bmatrix} -1/2 & -1 \\ -2 & -6 \end{bmatrix} \begin{bmatrix} V_{c_1} \\ V_{c_2} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 2 & -4 & -2 \end{bmatrix} \begin{bmatrix} E \\ \dot{E} \\ \ddot{E} \end{bmatrix} \quad . \end{aligned} \tag{26}$$

Again reducing, the state equations are now obtained in final form:

$$\begin{bmatrix} \dot{V}_{c_1} \\ \dot{V}_{c_2} \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ -1/2 & -1 \end{bmatrix} \begin{bmatrix} V_{c_1} \\ V_{c_2} \end{bmatrix} + \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} E \\ \dot{E} \\ \ddot{E} \end{bmatrix} \quad . \tag{27}$$

The output equations become:

$$\begin{bmatrix} I_{in} \\ V_{out} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -1/5 & 4/5 \end{bmatrix} \begin{bmatrix} V_{c_1} \\ V_{c_2} \end{bmatrix} + \begin{bmatrix} 2 & 2 & 1 \\ 1/5 & 2/5 & 0 \end{bmatrix} \begin{bmatrix} E \\ \dot{E} \\ \ddot{E} \end{bmatrix} \quad . \tag{28}$$

The input current in this circuit is independent of the state, and we may immediately write:

$$I_{in} = \ddot{E} + 2\ \dot{E} + 2\ E \quad , \tag{29}$$

or

$$Y_{in}(s) = s^2 + 2s + 2 \quad . \tag{30}$$

In Section C, the possibility of singular matrices twice arose. It may now be seen that the reduction algorithm just described is applicable in both cases to remove dependent elements from the state vector.

E. REMARKS

It should be clear from the previous discussion that this procedure will in theory yield the differential equations of any linear network possessing a solution. Dependent equations, the theoretical source of difficulty, are detected by the reduction of matrices to row echelon form and subsequent reduction of the dimension of the state vector if required. The numerical difficulties of this scheme, however, demand further investigation. Although much has been written on the numerical solution of linear algebraic equations, the situation here is more complex than usual, since errors can propagate over more than one reduction.

Once the state equations are determined, an analysis may proceed in several directions. If transfer functions and frequency responses are desired, the so-called Souriau-Frame algorithm may be employed to obtain the matrix $[sI - A]^{-1}$, the transformed state transition matrix.[6,7] This recursive method, requiring at each step only the multiplication of two matrices and the addition of a trace, seems to be in some disrepute. Upon closer inspection, the difficulty seems to be its poor efficiency (requiring $\sim n^4$ multiplications) when used to find the characteristic polynomial of a matrix as compared to other methods (requiring $\sim n^3$ multiplications). Here, however, we need an entire matrix of rational functions, not just the characteristic polynomial of A. Once $\Phi(s)$ is found, the matrix of transfer functions desired follows from

$$Y(s) = \{C\Phi(s)\,[B+B_1 s+\ldots+B_m s^m]+D+D_1 s+\ldots D_m s^m\}\,U(s)\,, \quad (31)$$

where B_i and D_i are matrices associated with the i^{th} derivatives of input quantities. Notice that the calculation of many different transfer functions for one network requires little more computation than for one alone. Although time responses may be obtained from this relation by the usual technique of factoring the characteristic polynomial and performing a partial fraction expansion of each entry in the transfer-function matrix, a more accurate and efficient method would seem to be a Taylor-series approximation to the state transition matrix exp (At).[8] This approach is open only to those analysis methods generating state equations.

A complete computer program implementing the details of this paper has been written both in FORTRAN II for the 7094 and FORTRAN-63 for the 1604. The input to the program describes the network in terms of elements and their interconnections; the output is presently the A, B, C, and D matrices and a matrix of desired transfer functions. These transfer functions are left in memory in a way that facilitates their use by any of the standard routines for manipulating rational functions. Unfortunately, information on accuracy and running time is not yet available, but these should be as good or better than passive network analysis routines using node or loop techniques.

The form of the equations derived using this method indicate two obvious extensions. The first is to network analysis involving the variation of several element values in the course of one run. If these elements are removed from the network and a new input and output defined for each, they may be added after the state equations have been found. For frequency response analysis, this approach is very much akin to a hybrid analysis proposed by So.[12] A second and equally interesting extension is to nonlinear networks composed chiefly of linear elements. The nonlinear elements are withheld from the network as outlined above and are added after the state equations of the linear portion have been obtained. The result would be the generation of a set of first-order nonlinear differential equations in exactly the form required by most numerical differential equation solving routines.

F. REFERENCES

1. T. R. Bashkow, "The A Matrix, New Network Description," IEEE Trans., CT-4 (September 1957), pp. 117-119.
2. P. R. Bryant, "The Order of Complexity of Electrical Networks," Proc. IEE, 106C (June 1959), pp. 174-188.
3. P. R. Bryant, "The Explicit Form of Bashkow's A Matrix," IEEE Trans., CT-9 (September 1962), pp. 303-306.
4. H. E. Koenig and Y. Tokad, "State Models of Systems of Multi-Terminal Linear Components," IEEE Convention Record (April 1964).
5. A. Dervisoglu, "Bashkow's A Matrix for Active RLC Networks," IEEE Trans., CT-11 (September 1964), pp. 404-407.
6. L. A. Zadeh and C. A. DeSoer, Linear System Theory: The State Space Approach, New York: McGraw-Hill, 1963.
7. J. S. Frame, "Matrix Functions and Applications, Part IV: Matrix Functions and Constituent Matrices," IEEE Spectrum, 1 (June 1964), pp. 123-131.
8. M. L. Liou, "A Numerical Solution of Linear Time-Invariant Systems," Proc. Third Allerton Conf., Urbana, Ill., October 1965.
9. A. Dervisoglu, "State Models of Active RLC Networks," Report R-237, Coordinated Science Lab., Univ. of Ill., December 1964.
10. S. K. Mitra, "A Unique Synthesis Method of Transformerless Active RC Networks," J. Franklin Inst., 274 (August 1962), pp. 115-129.
11. H. C. So, "Computational Aspects of the Inversion of a Tree Matrix," IEEE Trans., CT-12 (June 1965), pp. 271-272.
12. H. C. So, "On the Hybrid Description of a Linear n-port Resulting from the Extraction of Arbitrarily Specified Elements," to be published in IEEE Trans., CT-12 (December 1965).

A NUMERICAL SOLUTION OF LINEAR TIME-INVARIANT SYSTEMS

M. L. LIOU
Bell Telephone Laboratories, Incorporated
North Andover, Massachusetts

1. INTRODUCTION

Numerical solutions of ordinary differential equations have been considered by many people. The Runge-Kutta method and the various forms of predictor-corrector methods[1,2] have been used widely. They can solve non-linear as well as linear differential equations. Since a large number of systems can be described by a set of linear differential equations with constant coefficients (in the form of linear state space equations with constant matrices), it seems justified to develop a method which yields solutions faster than that of the general methods. Rane[3] did this in 1962. He approximated the function by a truncated Taylor series and used the forward differences to approximate the derivatives. His recursive formula requires less calculation and bookkeeping than the other general formulas. However, he failed to discuss the error introduced by the forward differences approximation. In this paper, a different recursive formula is derived together with the error expression. It requires even less calculation than that of Rane's under similar conditions.

2. PRELIMINARY ANALYSIS

Consider a system of linear differential equations with constant coefficients

$$\dot{\underline{x}}(t) = A\underline{x}(t) + B\underline{w}(t), \quad \underline{x}(0) = \underline{x}_0 \qquad (1)$$

where $\underline{x}(t)$ and $\underline{w}(t)$ are respectively mx1 and rx1 vector functions while A and B are respectively mxm and mxr constant matrices.

The exact solution of Equation (1) is [4]

$$\underline{x}(t) = e^{At}\,\underline{x}_0 + e^{At}\int_0^t e^{-A\tau}\,B\underline{w}(\tau)\,d\tau \qquad (2)$$

where

$$e^{At} = \sum_{k=o}^{\infty} \frac{A^k t^k}{k!}\,,\; A^o = I = \text{identity matrix} \qquad (3)$$

The matrix series given in the above equation can be shown to be uniformly convergent in any finite interval.[4]

When $t = nT$, from Equation (2),

$$\underline{x}(nT) = e^{AnT}\,\underline{x}_o + e^{AnT}\int_o^{nT} e^{-A\tau}\,B\underline{w}(\tau)d\tau \qquad (4)$$

Similarly, when $t = (n+1)T$

$$\underline{x}[(n+1)T] = e^{A(n+1)T}\,\underline{x}_o + e^{A(n+1)T}\int_o^{(n+1)T} e^{-A\tau}\,Bw(\tau)d\tau \qquad (5)$$

Multiplying Equation (4) by e^{AT} and subtracting it from Equation (5) gives

$$\underline{x}[(n+1)T] = e^{AT}\,\underline{x}(nT) + e^{A(n+1)T}\int_{nT}^{(n+1)T} e^{-A\tau}\,B\underline{w}(\tau)d\tau \qquad (6)$$

A recursive formula can be derived from the above equation.

3. THE RECURSIVE FORMULA

When the forcing function $\underline{w}(t)$ is either constant or consisting of impulses during the interval $nT \leq t < (n+1)T$, the integral in Equation (6) can be solved exactly and the corresponding recursive formulas can be derived accordingly.[5] The recursive formula for piecewise constant forcing functions can also be used to obtain approximate solutions when the forcing function is an arbitrarily continuous function over the entire range.[5] However, the approximation is poor. A much better approximation can be obtained by applying Simpson's rule[1] to the integral in Equation (6). This gives

$$x[(n+1)T] = e^{AT}[x(nT) + \frac{T}{6} B\underline{w}(nT)] + \frac{2T}{3} e^{A\frac{T}{2}} B\underline{w}[(n+\tfrac{1}{2})T]$$

$$+ \frac{T}{6} B\underline{w}\,[(n+1)T] + \underline{E} \qquad (7)$$

where the error term $\underline{E}$ can be expressed as

$$\underline{E} = - \frac{T^5}{2880} e^{A[(n+1)T-\zeta]} [B\underline{w}''''(\zeta) - 4AB\underline{w}'''(\zeta)$$

$$+ 6A^2B\underline{w}''(\zeta) - 4A^3B\underline{w}'(\zeta) + A^4B\underline{w}(\zeta)], \quad nT<\zeta<(n+1)T \qquad (8)$$

It should be noted that the error expression is invalid if the derivatives of $\underline{w}$ in Equation (8) do not exist. When T is sufficiently small, the error term in Equation (7) can be neglected.

Let $$e^{A\frac{T}{2}} = M \qquad (9)$$

Neglecting the error term, the recursive formula of Equation (7) becomes

$$x[(n+1)T] \cong M^2 [\underline{x}(nT) + \frac{T}{6} B\underline{w}(nT)] + \frac{2T}{3} MB\underline{w}\,[(n + \tfrac{1}{2})T]$$

$$+ \frac{T}{6} B\underline{w}[(n+1)T] \qquad (10)$$

The matrices M^2, $\frac{T}{6}$ B, and $\frac{2T}{3}$ MB can be evaluated once and for all at the beginning. The above equation requires $m^2 + 3$ mr multiplications and $m^2 + 3$ mr - m additions at each step. The amount of calculation at each step is the same as compared with Rane's method.[3]

4. EVALUATION OF MATRIX M

The matrix $e^{A\frac{T}{2}}$ may be determined by use of the Sylvester expansion theorem, by means of inverse Laplace transformation, or by inspection of the state-variable diagram.[6] However, all of these methods are impractical when the dimension of the system is large and the problem is to be solved by digital computers. For a given time increment T, the matrix $e^{A(T/2)}$ can be expressed by an infinite matrix series. For the numerical purpose, this series has to be truncated to obtain an approximate matrix. The number of terms required for computation can be determined as follows. Let the norm of matrix A be

$$\|A\| = \sum_{i,j=1}^{m} |a_{ij}| \tag{11}$$

then it can be shown that

$$\|A^k\| \leq \|A\|^k, \; k=1,2,\ldots \tag{12}$$

From Equations (9) and (3)

$$M = e^{A\frac{T}{2}} = \sum_{k=o}^{\infty} \frac{A^k \left(\frac{T}{2}\right)^k}{k!} \quad , \; A^o = I = \text{identity matrix} \tag{13}$$

This matrix series is uniformly convergent.

The norm of each term in the above matrix series is

$$\left\| \frac{A^k\left(\frac{T}{2}\right)^k}{k!} \right\| = \frac{\left(\frac{T}{2}\right)^k}{k!} \|A^k\| \leq \frac{\left(\frac{T}{2}\right)^k}{k!} \|A\|^k, \; k=1,2,\ldots \tag{14}$$

The ratio of the kth term to the (k-1)th term in the right hand side of the above equation is

$$\frac{\dfrac{\left(\frac{T}{2}\right)^k}{k!} \|A\|^k}{\dfrac{\left(\frac{T}{2}\right)^{k-1}}{(k-1)!} \|A\|^{k-1}} = \frac{\frac{T}{2} \|A\|}{k} \tag{15}$$

Set this ratio to within a prescribed value ϵ, then

$$k \geq \frac{T}{2\epsilon} \|A\| \tag{16}$$

The number of terms required in numerical calculation of M [Equation (13)] for given values of T and ϵ may be determined by Equation (16). Due to the fact that the value of ϵ is hard to specify, an iterative procedure is developed below.

Let the smallest value of k determined by Equation (16) be K, then

$$M = \sum_{k=o}^{K} \frac{A^k\left(\frac{T}{2}\right)^k}{k!} + R \tag{17}$$

Where the remainder matrix

$$R = \sum_{k=K+1}^{\infty} \frac{A^k\left(\frac{T}{2}\right)^k}{k!} \tag{18}$$

By the definition of the norm of a matrix, every element of the matrix A^k is less than or equal to $\|A^k\|$ and hence $\|A\|^k$.

It follows that each element in the remainder matrix, r_{ij}, can be expressed by the following inequality

$$|r_{ij}| \leq \sum_{k=K+1}^{\infty} \frac{\|A\|^k \left(\frac{T}{2}\right)^k}{k!} \tag{19}$$

From Equation (16)

$$\epsilon \geq \frac{\|A\| \frac{T}{2}}{k} , \ k \geq K \tag{20}$$

Substituting this relation into Equation (19) gives

$$|r_{ij}| \leq \frac{\|A\|^{K+1} \left(\frac{T}{2}\right)^{K+1}}{(K+1)!} \quad [1+\epsilon+\epsilon^2+\ldots]$$

$$= \frac{\left(\frac{T}{2}\ \|A\|\right)^{K+1}}{(K+1)!} \quad \frac{1}{1-\epsilon} \tag{21}$$

let

$$\tilde{M} = \sum_{k=o}^{K} \frac{A^k \left(\frac{T}{2}\right)^k}{k!} \tag{22}$$

If each element in matrix M is required to within, at least, d significant digits accuracy, then

$$|r_{ij}| \leq 10^{-d} \ |\tilde{m}_{ij}| \tag{23}$$

where $\tilde{m}_{ij}$ is an element of matrix $\tilde{M}$.

The computational sequence of determining the number of terms, can be summarized as follows:

(i) Give an initial value ϵ.

(ii) Find the smallest integer K from Equation (16) (T and A are given).

(iii) Calculate $\tilde{M}$ from Equation (22).

(iv) Find the upper bound of $|r_{ij}|$ by Equation (21).

(v) Compare each element value $\tilde{m}_{ij}$ obtained in (iii) with the value obtained in (iv). If Equation (23) is not satisfied (d is given), decrease the value of ϵ and repeat (i). Otherwise, the sequence is ended and the matrix $\tilde{M}$ calculated in (iii) is the desired approximation of M.

5. CONCLUSION

A numerical solution of linear state space equations using Simpson's rule has been obtained. The transition matrix $e^{A(T/2)}$ is approximated by a truncated matrix series. This method yields solutions faster than that of other methods.

A brief comparison of this method and Rane's method[3] is given below. When K = 4, which is equivalent to the fourth degree accuracy of Rane's method, the matrices M^2, $\frac{T}{6}$ B, and $\frac{2T}{3}$ MB of this method require much less calculation and bookkeeping than that of Rane's. The recursive formula of this paper is similar to that of Rane's; which requires the same amount of calculation at each step. However, Rane's method does not give the error expression due to the forward differences approximation.

REFERENCES

1. F. B. Hildebrand, "Introduction to Numerical Analysis," McGraw-Hill Book Company, New York, 1956.

2. R. W. Hamming, "Numerical Methods for Scientists and Engineers," McGraw-Hill Book Company, New York, 1962.

3. D. S. Rane, "On the Numerical Solution of Linear and Nonlinear System Models," Ph. D. Thesis, Michigan State University, East Lansing, Michigan, 1962.

4. R. Bellman, "Stability Theory of Differential Equations," McGraw-Hill Book Company, New York, 1953.

5. J. J. Lang and Mrs. J. M. Littlefield, private communication.

6. J. T. Tou, "Modern Control Theory," McGraw-Hill Book Company, New York, 1964.

PRACTICAL FILTER DESIGN USING MATHEMATICAL OPTIMIZATION*

A. D. WAREN and L. S. LASDON
Electronic Research Division, Clevite Corporation, Cleveland, Ohio
Division of Organizational Science & Systems Research Center,
Case Institute of Technology, Cleveland, Ohio

ABSTRACT

This paper proposes a solution to the problem of designing a filter of given structure, incorporating non-ideal elements, to meet or exceed given insertion loss specifications subject to element value bounds. This problem is reformulated as a nonlinear programming problem, i. e., minimize an objective function subject to inequality constraints, whose solution yields a filter optimal in a Chebychev sense.

To solve this problem, a recent penalty function approach is used, which converts the constrained problem into a sequence of unconstrained minimizations. These minimizations are carried out using a recent, very efficient, descent technique. The overall method is especially amenable to computer implementation.

These techniques have been applied to the computer design of cascade crystal-realizable lattice filters. Some results are presented, and these are uniformly good.

I. INTRODUCTION

Mathematical programming techniques have been widely used in Operations Research, Economics, and related fields. However, the use of these techniques for the computer design of engineering systems is an area which is thus far relatively unexplored, although some work has been carried out in the chemical[1] and structural[2, 3] areas. This paper presents an approach to the design by digital computer of electrical filters which appears to have generalizations to other areas as well.

The classical statement of the network synthesis problem is as follows: Given an input and response function, determine a network which will produce the given response when excited by the specified input, or determine that no such network exists. For the design of practical networks, however, there are many additional factors which must be considered and indeed the statement of the practical design problem generally bears little resemblence to the statement of the classical synthesis problem. In fact it is this difference in the formulation of the two problems that leads to the existence of a second major area in network synthesis, the "approximation problem."

Moreover, existing synthesis techniques do not enable the user to specify in advance bounds on allowable element values. Obviously

---

[*] This work was supported in part by the U. S. Army Electronics Laboratory, Fort Monmouth, New Jersey.

such bounds exist and indeed are often the reason that theoretically designed networks are not satisfactory from the practical point of view. Further, the theoretical synthesis techniques assume the existence of ideal elements. Even simple dissipative effects are difficult to account for except in the most elementary cases,[4, 5] and stray capacitances are almost impossible to incorporate.

In addition, in the practical filter design problem the network topology is often specified in advance. Again, except for some simple cases,[6, 7] this type of problem cannot be satisfactorily handled by existing techniques.

As an example of the above restrictions, consider the following problem. Design a single sideband (lower side band) filter at 10.5 Mc to have a 1 db bandwidth of 3. 0 kc and a 60 db bandwidth of less than 6. 0 kc. The steep side 60:3 db skirt ratio is to be less than 1. 4. The resulting network is to consist of the cascade connection of four quartz resonator half lattice sections with two quartz crystals per section. Shunt capacitances may be included across the lattice arms along with tuned circuits at the input and output of each half lattice. The pass band ripple must not exceed 1 db, the stopband rejection must be greater than 70 db, and the entire structure must be realizable with readily available components.

At the moment, there exists no synthesis technique capable of solving the above problem. Similar but less stringent designs are generally handled on the following basis. Obtain an ideal lossless design which exceeds the given specifications. Incorporate losses and stray effects, juggle element values to obtain realizable figures, and examine the response of the resulting structure. If the result is satisfactory, fine, if not try again.

Because of the above mentioned problems, the concept of using a computer iterative technique has often been proposed as a logical solution.[8, 9, 10, 11, 12, 13, 14] Several approaches have been tried but, one must conclude, with little success. Successful results have been achieved in solving the approximation problem[15, 16, 17, 18] by iterative techniques but the resulting network still has all the shortcomings of any present-day synthesis scheme. Let us briefly examine the possible reasons for the "failure" of this conceptually powerful approach in solving the problem of network design.

The first difficulty arises in the formulation of the problem to be solved by the computer. The desired network has generally been required to minimize the squared error between its network function and some specified function.[8, 9, 10, 11, 12] The sample problem quoted earlier does not lend itself naturally to such a formulation -- it asks only that the network function be above or below certain levels in certain ranges of the frequency band. Further, the squared-error criterion has been found difficult to minimize. In addition, little or no attention has been paid to element value restrictions.

Secondly, the optimization techniques which have been used to reduce the error criterion have been at best, elementary e. g. , first order gradient techniques, direct variation of parameters, or Newton Raphson schemes.

In seeking to avoid the above difficulties, the problem is here formulated essentially as given to the designer, i. e., choose the (bounded) element values in a fixed structure so that the insertion loss lies above or below certain levels within given frequency bands. This formulation transforms the filter design problem into the problem of finding a set of element values satisfying a finite number of inequalities (Section II). This problem is, in turn, converted into a nonlinear programming problem, i. e., minimize some objective function subject to a system of inequality constrains (Section III).

The method selected for solving this problem is that of Fiacco and McCormick,[19] which reduces the above problem to a sequence of unconstrained minimization problems. This method appears to be the best now available for the case of nonlinear constraints. It is one of the few for which computational experience is reported, and that experience is quite good.

The computational properties of the method depend on how fast and accurate is the technique for solving the unconstrained "subproblems". A recent method due to Fletcher and Powell,[20] which is a "second order" gradient technique has much better convergence properties than first order methods, and is used for the unconstrained minimization. Further, in the Fletcher-Powell approach, the objective function must be minimized along a given direction with respect to a step-size parameter. Here, cubic interpolation[21] combined with a sophisticated step size selection technique is used with good success. All these methods are discussed in Section III.

The union of this new formulation of the problem with what are considered to be the best available methods has yielded excellent results (Section IV). Several filters have been designed including the single side band filter mentioned at the beginning of the paper. This has also been fabricated and represents an advance in the present state of the art.

II. PROBLEM FORMULATION

1. Inequalities

The filter design problem is viewed here as the determination of a set of element values for a given network structure which a) are within given bounds and, b) yield an insertion loss function meeting given specifications. These specifications require the insertion loss to exceed or equal given values within stopband regions and to be less than or equal to given levels in the pass band region.

In order that only a finite number of insertion loss constraints (b above) need be considered, the insertion loss specifications are assigned at a finite set of frequencies, rather than over an entire frequency band. With proper choice of these frequencies, the continuity properties of the insertion loss function insure that if the specifications are met at the assigned frequencies then they will be met everywhere.

Let $$F(\underline{e}, \omega_i) \quad F_i(\underline{e}) \qquad (1)$$

be the value of the insertion loss function at frequency ω_i for element values $\underline{e} = (e_1, \ldots, e_n)^T$. Then the problem is to determine $\underline{e}$ satisfying the following system of inequalities:

$$F_i(\underline{e}) \geq \alpha_i \qquad i=1, \ldots, a, b+1, \ldots, c$$

$$F_i(\underline{e}) \leq \alpha_i \qquad i=a+1, \ldots, b \qquad (2)$$

and

$$\ell_i \leq e_i \leq u_i \qquad i=1, \ldots, n \qquad (3)$$

The frequencies $\omega_i, \ldots, \omega_a, \omega_{b+1}, \ldots, \omega_c$ are the stopband frequencies, $\omega_{a+1}, \ldots, \omega_b$ are the passband frequencies, the α_i are the insertion loss specifications and the ℓ_i and u_i are lower and upper bounds, respectively for the i^{th} element value, e_i.

2. ξ Formulation

There may of course be many vectors $\underline{e}$ satisfying the constraints (1), (2) (i. e., many satisfactory filters) or there may be none at all. In either case a "best filter" can be meaningfully defined as one which exceeds the specifications at each point by the maximum amount or fails to meet the specifications by the minimum amount possible. The solution to the following nonlinear programming problem represents such a filter.[23]

Determine $(\underline{e}^T, \xi)$ such that

$$\xi \longrightarrow \text{minimum} \qquad (4)$$

subject to

$$F_i(\underline{e}) + \theta_i \xi \geq \alpha_i \qquad i=1, \ldots, a, b+1, \ldots, c$$

$$F_i(\underline{e}) - \theta_i \xi \leq \alpha_i \qquad i=a+1, \ldots, b \qquad (5)$$

and

$$\ell_i \leq e_i \leq u_i \qquad i=1, \ldots, n \qquad (6)$$

Here ξ is an additional independent variable and the θ_i are non-negative constants.

Initially, consider the case when all θ_i are unity. Then for any vector $(\underline{e}^T, \xi)$ satisfying the inequalities (5), (6) with $\xi < 0$, the insertion loss specifications are met or exceeded at all specified frequencies by at least $|\xi|$ db (see Fig. 1). If $\xi < 0$ and at least one of the constraints (5) is binding (satisfied as an equality) then $|\xi|$ is the minimum amount by which the specifications are exceeded. If $\xi > 0$ and at least one of the constraints is binding then ξ is the maximum amount by which the specifications fail to be satisfied (see Fig. 2).

At a minimum of the problem represented by (4), (5), and (6) at least one of the constraints (5) must be binding, otherwise ξ could be reduced without violating any of the constraints. Thus if min $\xi < 0$, the solution to (4)-(6) defines an optimal filter which maximizes the minimum amount by which the insertion loss specifications are met. Similarly if min $\xi > 0$ then the specifications cannot be met at all given frequencies and the resulting filter minimizes the maximum amount by which the specifications are not met. Of course in either case the element value constraints are satisfied.

Relaxing the requirement that all θ_i be unity and requiring only that θ_i be non-negative (in Eq. 5) allows great flexibility in "curve shaping". For example, given an initial design in which the stop band rejection is satisfactory but the passband ripple too great, then solving (4)-(6) with

$$\theta_i = 0 \qquad i=1, \ldots, a, b+1, \ldots, c.$$

$$\theta_i = 1 \qquad i=a+1, \ldots, b$$

will generate a design which minimizes the ripple and maintains the present stopband level. Values for the θ_i other than 0 or 1 may also be used to advantage.

III. SOLUTION TECHNIQUE

1. Introduction

Although formulating practical filter design as a nonlinear programming problem is attractive, the matter of solving this problem still remains. Only in the last few years have solution techniques for such problems developed sufficiently to make this approach feasible.

The chief difficulty in solving (4)-(6) is due to the constraints (5). For a filter of reasonable size the functions $F_i(\underline{e})$ are highly nonlinear. Thus a solution technique which works well on problems with markedly nonlinear constraints is needed. It is also desirable that the technique decrease ξ monotonically, while maintaining all constraints satisfied, since this gives an improved filter at each iteration. Furthermore the method should be able to solve problems which may have local minima i. e., problems that are not convex. This is the case for (4)-(6). A recent technique developed by Fiacco and McCormick[19, 24, 24] has these properties and additional advantages. This method and associated solution techniques are now described.

2. The Sequential Unconstrained Minimization Technique

The sequential unconstrained minimization technique of Fiacco and McCormick solves an inequality constrained minimization problem by converting it to a sequence of unconstrained minimization problems. The constrained problem may be formulated as

$$\underset{\underline{x}}{\text{minimize}}\ f(\underline{x}) \tag{7}$$

$$\text{Subject to} \quad g_i(\underline{x}) \geq 0\,, \qquad i=1, \ldots, m \tag{8}$$

The problem (4)-(6) is easily converted to this form.

Define the function:

$$P(\underline{x}, r) = f(\underline{x} + r \sum_{i=1}^{m} \frac{1}{g_i(\underline{x})} \tag{9}$$

where $r > 0$. Let

$$R = \left\{\underline{x} \mid g_i(\underline{x}) \geq 0\right\}, \quad i=1, \ldots, m \tag{10}$$

and

$$R^o = \left\{\underline{x} \mid g_i(\underline{x}) > 0\right\}, \quad i=1, \ldots, m \tag{11}$$

The set R is called the constraint set and R^o is its interior. Choose $r_1 > 0$, $\underline{x}_o \in R^o$ and consider the problem of minimizing $P(\underline{x}, r_1)$ starting from $\underline{x}_o$ subject to no constraints. Intuitively one would expect that a minimum of $P(\underline{x}, r_1)$ exists in R^o since on the boundary of this set, i. e., for some $g_i(\underline{x}) = 0$, $P(\underline{x}, r_1) \longrightarrow +\infty$. Thus the trajectory of steepest descent leading from $\underline{x}_o$, a path on which $P(\underline{x}, r_1)$ is strictly decreasing, cannot penetrate the boundary of the constraint set. The minimizing point depends, of course, on the choice of r_1, and is denoted by $\underline{x}(r_1)$. By the above arguments, $\underline{x}(r_1) \in R^o$.

Now repeat this minimizing process for a sequence of r values $r_1 > r_2 > \ldots, r_k > 0$. Each minimizing point $\underline{x}(r_k)$ will be in R^o. Further, by reducing r the influence of the term $r \sum_{i=1}^{m} 1/g_i(\underline{x})$, which "penalizes" closeness to the constraint boundaries, is reduced and in minimizing P more effort is concentrated upon reducing $f(\underline{x})$. Thus the sequence of points $\underline{x}(r_1)$, $\underline{x}(r_2)$, . . . , $\underline{x}(r_k)$ can come closer and closer to the boundary of the set R if it is profitable, in terms of reducing f, to do so. One would thus expect that as r approaches zero, the minimizing point $\underline{x}(r)$ approaches the solution of the original problem (7), (8). Such a method is particularly attractive when dealing with problems with markedly nonlinear constraints since it approaches the solution value from inside the constraint set. Motion along the boundaries of this set, which can be very cumbersome when the boundaries have large curvature, is completely avoided.

Fiacco and McCormick prove that, under certain regularity and convexity assumptions, as $r_k \longrightarrow 0$ the values $P\left[\underline{x}(r_1), r_1\right]$, . . . , $P\left[\underline{x}(r_k), r_k\right]$, . . . , approach v_o, the minimum value of $f(\underline{x})$ subject to (8). In addition, it is shown that f is monotonically non-increasing as r is reduced.[19, 24, 25]

3. The Fletcher Powell Minimization Method

The inequality constrained minimization problem has been converted to a sequence of unconstrained P-function minimizations. Now, the question of efficiently minimizing the P-function arises. The insertion loss functions $F_i(\underline{e})$ are complex nonlinear functions of the element values $\underline{e}$. Thus the P-function for this application is very complex and markedly nonlinear, hence some iterative technique must be used for the minimization.

Since the gradients of the functions $F_i(\underline{e})$ can be calculated analytically (Section IV), any of the gradient descent schemes are feasible. However, much evidence exists[19, 20, 24] that first order gradient techniques will converge very slowly, if at all, for such complex functions. Alternative methods requiring second partial derivatives[19, 24] have weak convergence properties and would require second derivatives of the insertion loss, making them computationally infeasible for this application.

The method adopted is a recent scheme developed by Fletcher and Powell.[20] Only first derivatives of the function to be minimized are required, yet the technique has "second order" convergence i. e., it minimizes a positive definite quadratic form in n variables in n iterations. This method has found wide acceptance and is generally regarded as "the most powerful general procedure for finding a local minimum which is known at the present time."[20]

Central to the method is a symmetric positive definite matrix H_i, which is updated at each iteration, and which determines the current direction of motion, $\underline{s}_i$, by multiplying the current gradient vector. An iteration is described as follows:

$$H_o = \text{any positive definite matrix}$$

$$\underline{s}_i = -H_i \nabla f(\underline{x}_i) \tag{12}$$

Choose $\alpha_i = \alpha_i^*$ by minimizing $f(\underline{x}_i + \alpha_i \underline{s}_i)$

$$\underline{\sigma}_i = \alpha_i^* \underline{s}_i \tag{13}$$

$$\underline{x}_{i+1} = \underline{x}_i + \underline{\sigma}_i \tag{14}$$

$$H_{i+1} = H_i + A_i + B_i \tag{15}$$

where the matrices A_i and B_i are defined by

$$A_i = \frac{\underline{\sigma}_i \underline{\sigma}_i^T}{\underline{\sigma}_i^T \underline{y}_i}, \quad \underline{y}_i = \nabla f(\underline{x}_{i+1}) - \nabla f(\underline{x}_i) \tag{16}$$

$$B_i = \frac{-H_i \underline{y}_i \underline{y}_i^T H_i}{\underline{y}_i^T H_i \underline{y}_i} \tag{17}$$

Note that the numerators of A_i and B_i are both matrices, while the denominators are scalars.

Fletcher and Powell prove the following:

a) The matrix H_i is positive definite for all i. As a consequence of this, the method will always converge since

$$\frac{d}{d\alpha_i} f(\underline{x}_i + \alpha_i \underline{s}_i)_{\alpha_i = 0} = -\nabla f^T(\underline{x}_i) H_i \nabla f(\underline{x}_i) < 0 \tag{18}$$

i. e., the function is initially decreasing along the direction $\underline{s}_i$, so that the function can be decreased at each iteration by minimizing down $\underline{s}_i$.

b) When the method is applied to the quadratic

$$f(\underline{x}) = \underline{c}^T \underline{x} + \underline{x}^T A \underline{x} \tag{19}$$

with A positive definite:

(i) The minimum is reached in n steps.

(ii) The matrix H_i converges to the inverse of the matrix of second partials of the quadratic after n iterations, i. e.,

$$H_n = A^{-1} . \tag{20}$$

When applied to a general function, H_i tends to the inverse of matrix of second partials of the function evaluated at the minimum, since as the minimum is approached the second order terms in the Taylor series expansion predominate. Numerical results are available[20] illustrating the rapid convergence of this method and its superiority to the method of steepest descents.

4. One Dimensional Minimization

Each iteration of the Fletcher-Powell method requires minimization of the one-dimensional function $f(\underline{x}_i + \alpha_i \underline{s}_i)$. The efficiency of the entire procedure depends critically upon the efficiency of this single variable search. A modification of a technique described by Fletcher and Reeves[21] is used here. This method solves the problem of finding $\alpha_i = \alpha_i^*$ to minimize

$$g(\alpha_i) = f(\underline{x}_i + \alpha_i \underline{s}_i) \tag{21}$$

in three stages. It uses the derivative

$$g'(\alpha_i) \quad \frac{dg}{d\alpha_i} = \underline{s}_i^T \nabla f(\underline{x}_i + \alpha_i \underline{s}_i) \ . \tag{22}$$

The first stage calculates a step size for use by the second. The second stage establishes bounds on α_i^* and the third stage interpolates its value using cubic interpolation.[22]

IV. APPLICATION TO THE FILTER DESIGN PROBLEM

1. P-Function

Applying the Fiacco-McCormick technique to the problem (4)-(6) leads to the P-function

$$P(\underline{e}, \xi\ r) = \xi + r \left\{ \sum_{i=1}^{a} \frac{1}{F_i(\underline{e}) + \theta_i \xi - \alpha_i} + \sum_{i=a+1}^{b} \frac{1}{\alpha_i - F_i(\underline{e}) + \theta_i \xi} \right.$$

$$\left. + \sum_{i=b+1}^{c} \frac{1}{F_i(\underline{e}) + \theta_i \xi - \alpha_i} + \sum_{i=1}^{n} \frac{1}{u_i - e_i} + \sum_{i=1}^{n} \frac{1}{e_i - \ _i} \right\} . \tag{23}$$

Minimizing P using the Fletcher-Powell method requires the computation of the gradient of P with respect to e and ξ. This is easily obtained if the gradient of $F_i(\underline{e})$ is available for all i. In practice, the major portion of the computation time is spent in this gradient evaluation, and it is important that it be done efficiently. The form of this computation depends on the network topology. An example for cascade networks is discussed below.

2. Derivative Calculation

For a doubly terminated cascade filter structure, the insertion loss in db is

$$\alpha = 10 \log_{10} \left\{ |A|^2 \right\} + 10 \log_{10} \left\{ \frac{R_L}{4R_S} \right\}$$

The partial derivative of the insertion loss with respect to the j^{th} element of the i^{th} subnetwork is

$$\frac{\partial \alpha}{\partial e_{ij}} = \frac{20 \log_{10} e}{|A|^2} \left\{ \mathrm{Re}A \cdot \frac{\partial \mathrm{Re}A}{\partial e_{ij}} + \mathrm{Im}A \cdot \frac{\partial \mathrm{Im}A}{\partial e_{ij}} \right\}$$

if e_{ij} is neither the source nor load resistance.

The overall ABCD parameters of a cascade structure can be obtained by the matrix multiplication, in proper order, of the ABCD matrices of the individual sections, i. e.,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \prod_{i=0}^{n+1} \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix}$$

where the 0^{th} and $n+1^{st}$ ABCD parameters correspond to the input and output resistances. Moreover, it is possible to obtain the partial derivatives of the overall ABCD parameters with respect to e_{ij} in a relatively simple fashion using matrix differentiation. This requires the differentiation of only one submatrix for each element, i. e.,

$$\begin{bmatrix} \frac{\partial A}{\partial e_{ij}} & \frac{\partial B}{\partial e_{ij}} \\ \frac{\partial C}{\partial e_{ij}} & \frac{\partial D}{\partial e_{ij}} \end{bmatrix} = \frac{\partial}{\partial e_{ij}} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \prod_{k=0}^{i-1} \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix} \times \begin{bmatrix} \frac{\partial A_i}{\partial e_{ij}} & \frac{\partial B_i}{\partial e_{ij}} \\ \frac{\partial C_i}{\partial e_{ij}} & \frac{\partial D_i}{\partial e_{ij}} \end{bmatrix}$$

$$\times \prod_{k=i+1}^{n+1} \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix}$$

since the derivative of the k^{th} ABCD matrix with respect to e_{ij} is identically zero for $i \neq k$.

Therefore, in order to evaluate the insertion loss of a cascade structure, and the derivatives of the insertion loss with respect to all the elements, it is sufficient to determine the ABCD parameters of each section and their derivatives with respect to the elements contained in that section. Then a sequence of matrix multiplication (using complex numbers) will generate the required results. Since the insertion loss function and its derivatives are required at all specified frequencies, one such computation must be done at each frequency.

Perhaps the most important message of this report is that, despite some theoretical difficulties due to the non-convexity of $F_i(\underline{e})$, the methods presented have worked well on a number of significant filter design problems. Convergence has been rapid, and the final designs are excellent. A summary of results follows.

V. RESULTS

1. Filter Topology

The specific network structure selected to demonstrate the previously described techniques is the double terminated, cascade, half-lattice resonator filter with tuned circuits between half-lattice sections and at the input and output. In the actual computer program memory limitations require restriction of the network to one consisting of at most seven cascade half-lattice sections with no more than four resonators per lattice arm. Any filter satisfying these restrictions can be designed. Also, any elements can be arbitarily set to zero, providing considerable additional freedom in the selection of a specific filter topology.

2. Examples

The first example utilizing the previously described design techniques was a lossy four resonator filter consisting of two cascade half-lattice sections with one resonator per lattice arm. The design objectives required this filter to have a minimum of 30 db rejection over specified frequency ranges, and less than 4 db attenuation anywhere in the passband. Other requirements were a 30 db to 1 db skirt ratio of 1. 5 and a 1 db bandwidth of 10 kc at a 17. 6 Mc center frequency. The attenuation characteristics were specified at thirty frequencies, ten in the lower stopband, ten in the passband, and ten in the upper stopband. Realistic bounds were selected for all element values and Q's assigned to the resonators and tuned circuits.

Simple image parameter theory was used to give an initial feasible filter design (feasible in the sense that all element value constraints are satisfied). Figure 1 shows the insertion loss response for the initial design and, as is fairly obvious, this design is not particularly good. The final insertion loss characteristic (Fig. 2) displays equal ripple behavior in both pass and stopbands and is symmetric about the center frequency. This is indicative of an optimal design in the Chebychev sense.

For a second example, the single sideband specifications described in Section I were selected as design objectives. In this case it was considered desirable to have all resonator motional capacitances equal, and this equality constraint was incorporated in the program. The initial design was based on a modification of a design by Ho.[14] Figures 3 and 4 show the evolution of this particular filter.

3. Conclusions

Practical filter design has been formulated as a nonlinear programming problem and methods have been proposed for its solution. This formulation permits direct attack on the problem as given to the designer, circumventing the "approximation problem" inherent in classical techniques. Moreover realistic bounds can be assigned to all element values and losses and stray capacitance effects accounted for. The resulting filter is, in a sense meaningful to the design problem, optimal. The technique is quite flexible in that computation can be terminated when a satisfactory filter is attained rather than continuing to an optimal one. Also a curve-shaping option allows manipulation of portions of the insertion loss characteristic while other portions are maintained above or below specified levels.

The approach was programmed in Algol 60 for a Burroughs B5500 computer. Several filters have been designed with very satisfactory results. Computing time has been on the order of one hour per filter and it is felt that a significant reduction in this time can be obtained by using a more efficient linear search technique. This technique provides a practical and useful tool for network design.

ACKNOWLEDGMENT

The authors wish to thank G. P. McCormick and A. V. Fiacco for providing information concerning their penalty function approach to constrained minimization problems.

VI. REFERENCES

1. R. R. Hughes, E. Singer, and M. Souders, "Machine Design of Refineries," presented at the World Petroleum Congress, 1963.
2. L. A. Schmit, "Structural Design by Systematic Synthesis," Proc. 2nd National Conference Electronic Computation, Structural Division, Am. Soc. Civil Engineers, Pittsburgh, Pennsylvania, pp. 105-132, 1960.
3. R. L. Fox and L. A. Schmit, "Advances in the Integrated Approach to Structural Synthesis," AIAA 6th Annual Structures and Materials Conference, Palm Springs, California, pp. 50-66, April, 1965.
4. G. C. Temes, "First Order Estimation and Precorrection of Parasitic Loss Effects in Ladder Filters," I. R. E. Trans. on Circuit Theory, Vol. CT-9, pp. 385-399, December, 1962.
5. V. Belevitch, J. M. Goethals, and J. Neirynck, "Darlington Filters with Unequal Uniform Dissipation in Inductors and Capacitors," Philips Research Reports (Netherlands), Vol. 19, No. 5, pp. 441-468, October, 1964.
6. A. D. Waren and J. D. Schoeffler, "Approximation Problem for Resonator Ladder Filters," I. E. E. E. Trans. on Circuit Theory, Vol. CT-12, June, 1965.
7. T. R. O'Meara, "The Synthesis of a Class of Wide-Band Lattice Crystal Filters with Symmetrical Insertion Loss Characteristics," I. R. E. Trans. on Circuit Theory, Vol. CT-7, pp. 121-126, June, 1960.
8. M. R. Aaron, "The Use of Least Squares in System Design," I. R. E. Trans. on Circuit Theory, Vol. CT-3, pp. 224-231, December, 1965.
9. C. A. Desoer and S. K. Mitra, "Design of Lossy Ladder Filters by Digital Computer," I. R. E. Trans. on Circuit Theory, Vol. CT-8, pp. 192-201, September, 1961.
10. C. L. Semmelman, "Experience with a Steepest Descent Computer Program for Designing Delay Networks," Bell Telephone Laboratories, Internal Memo, May, 1962.
11. L. C. Peach and E. W. Weber, "Study of I. F. Transformers for Single Sideband Operation," I. I. T. Dept. of E. E., Addendum to the final report, U. S. Army Signal Supply Agency Contract No.

12. T. Murata, "The Use of Adaptive Constrained Descent in System Design," I. E. E. E. International Convention, Vol. 12, Pt. 1, March, 1964.

13. Y. Shigei, et al., "Network Design by the Taylor Series Method," Review of The Electrical Communication Laboratory, Vol. 10, pp. 483-500, September-October, 1962.

14. E. C. Ho, "S. S. B. and VHF Crystal Filters," 2nd, 3rd, and 4th Quarterly Reports, Signal Corps Contract No. DA 36-039 AMC-02305(E), September, 1963-June, 1964.

15. T. Fujisawa, "Theory and Procedure for Optimization of Low-Pass Attenuation Characteristics," I. E. E. E. Trans. on Circuit Theory, Vol. CT-11, pp. 449-456, December, 1964.

16. B. R. Smith and G. C. Temes, "An Iterative Approximation Procedure for Automatic Filter Synthesis," I. E. E. E. Trans. on Circuit Theory, Vol. CT-12, pp. 107-112, March, 1965.

17. J. D. Schoeffler, "A Solution to the Approximation and Realization Problems for Crystal Ladder Filters," I. E. E. E. International Convention Record, Pt. I, pp. 282-287, March, 1964.

18. T. Iedokoro, Y. Ishizaki and H. Watanabe, "On the Approximation and Synthesis for Tchebycheff Filters with Flat Group Delay Characteristics," Nippon Electric Co., Ltd., Monograph DEB 339, 1963.

19. A. V. Fiacco and G. P. McCormick, "Programming Under Nonlinear Constraints by Unconstrained Minimization: A Primal-Dual Method," Research Analysis Corporation, Technical Paper No. RAC-TP-96, September, 1963.

20. R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," British Computer Journal, pp. 1963-168, June, 1963.

21. R. Fletcher and C. M. Reeves, "Function Minimization by Conjugate Gradients," British Computer J., pp. 149-154, July, 1964.

22. W. C. Davidon, "Variable Metric Method for Minimization," A. E. C. Research and Development Report, ANL-5990, 1959.

23. G. Zoutendijk, Methods of Feasible Directions, American Elsevier Publishing Co., New York, pp. 66-68, 1960.

24. A. V. Fiacco and G. P. McCormick, "Computational Algorithm for the Sequential Unconstrained Minimization Technique for Nonlinear Programming," Management Science, Vol. 10, No. 4, pp. 601-617, 1964.

25. A. V. Fiacco and G. P. McCormick, "The Sequential Unconstrained Minimization Technique for Nonlinear Programming, A Primal-Dual Method," Management Science, Vol. 10, No. 2, pp. 360-366, 1964.

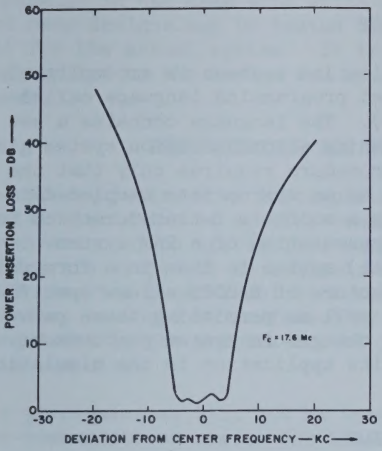


FIG. 1. INITIAL POINT FOR FIRST FILTER DESIGN

ERD 10860

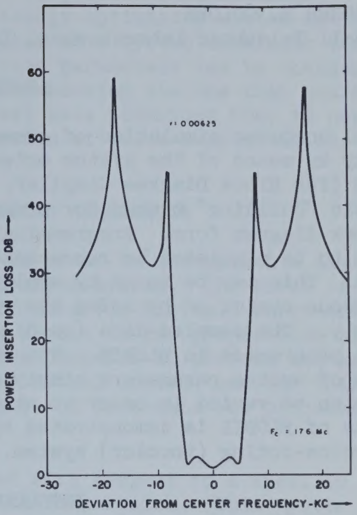


FIG. 2. FILTER No. I AFTER MINIMIZATION WITH LAST r.

ERD 10838

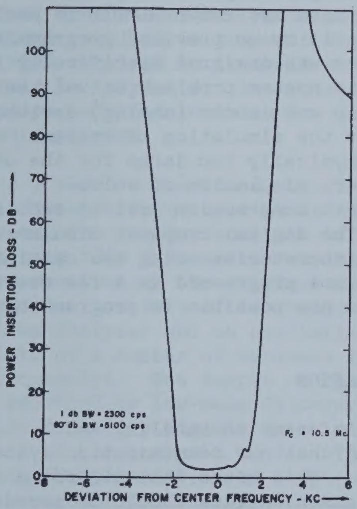


FIG. 3 INITIAL POINT FOR SECOND FILTER DESIGN

ERD 10840

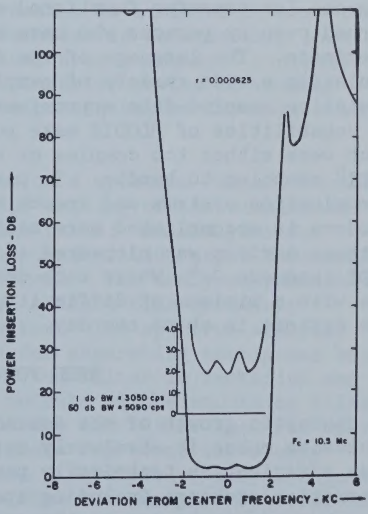


FIG. 4 SECOND FILTER AFTER MINIMIZATION WITH LAST r.

ERD 10843

DIGITAL COMPUTER SIMULATION OF COMMUNICATION SYSTEMS USING THE BLOCK DIAGRAM COMPILER: BLODIB

ROGER M. GOLDEN
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

ABSTRACT

Digital computer simulation of communication systems is accomplished readily by means of the system oriented programming language called BLODIB (for Block Diagram Compiler, B). The language contains a set of 30 basic "building" blocks for programming a communication system given in block diagram form. Programming procedure requires only that the system to be simulated be represented by an appropriate sampled-data system. This may be found by applying a suitable z-transformation to a continuous system or by using the representation of a PAM system directly. The sampled-data (or digital) system is then in a form which can be programmed in BLODIB. The structure of BLODIB allows specification of system parameters simply as well as permitting these parameters to be varied in order to study changes in system performance. The use of BLODIB is demonstrated by its application in the simulation of a voice-coding (vocoder) system.

INTRODUCTION

The study of complex communications systems by digital computer simulation is facilitated by use of the Block Diagram Compiler, BLODIB.[1] The programming language used in BLODIB is oriented towards communication system engineers and analysts who are already adept at handling block diagrams (or transfer functions) of complete systems. BLODIB is easily learned even by persons who have had little or no previous programming experience. The language of the compiler was designed specifically for simulating a wide variety of sampled-data system problems as well as simulating sampled-data approximations to continuous (analog) systems. The capabilities of BLODIB make possible the simulation of systems which were either too complex or were physically too large for the old BLODI[2] compiler to handle. In particular, simulation of voice-communication systems and speech bandwidth compression devices such as vocoders is accomplished more easily. The digital computer simulation of these devices was pioneered at Bell Laboratories using the original BLODI language.[3,4] While such devices were programmed in a few weeks time with a minimum of difficulty, it is now possible to program the same systems in about one day.

NEED FOR SIMULATION

The fantastic growth of the demand for information handling and processing makes it absolutely essential that our communication systems be as efficient as technically possible. This means that all forms of information exchange including speech communications should be encoded in such a way as to assure the most efficient use of a communication or data channel. The complexity of terminal equipment associated with such coding systems is such that a great deal of design, testing, and redesign must take place before these systems can be used on a regular basis. Once a given design is completed, the time element involved in testing and debugging (including redesign where necessary) such equipment introduces delays which are often costly. Many of these delays may be reduced or even avoided if the entire system is first evaluated by simulation on the digital computer. This technique provides the flexibility for systematically optimizing system design. The advantage of using a digital computer as a system simulator lies principally in the ease with which system parameters can be changed. Hence many designs may be tested before choosing the one that would be built for the actual system. It is just this technique that is now being applied to the design and evaluation of voice-communication systems.

PREREQUISITES FOR COMPUTER SIMULATION

Three prerequisites must be completed in order to accomplish a digital computer simulation of a voice-communication system. The first is the determination of an appropriate sampled-data (or digital) representation for the system to be studied; the second is the preparation of a BLODIB computer program which causes the computer to perform the same operations as would be performed by the actual system; and the third prerequisite is the preparation of input- or test-data for processing by the computer.

These prerequisites will now be treated with respect to a specific voice-communication system for bandwidth compression. This system is more commonly known as a vocoder (an acronym for voice-coder). The vocoder, invented over 35 years ago by Homer Dudley at Bell Telephone Laboratories, codes voices or speech signals for transmission over a communication channel. Parameters describing the speech are transmitted to a receiver where they are used to synthesize a replica of the original speech sounds. Bandwidth compression is achieved because the bandwidth required to transmit these parameters is considerably less than that required to transmit the original speech sounds. Probably, it would be helpful at this point to give a brief review of vocoder operation before describing how the above prerequisites for simulation are accomplished.

Figure 1 shows a functional block diagram of a spectrum channel vocoder and in itself can be used as an aid in the writing of the simulation program. The vocoder consists of two main parts: an analyzer for extracting that information which must be transmitted, and a synthesizer for reconstituting the original speech signal. The analyzer portion of the vocoder consists of two basic parts: a short-time frequency spectrum analyzer and an excitation detector. The spectrum analyzer consists of a number of bandpass filters for separating contiguous bands of frequencies. The output of each analyzing filter is rectified and then smoothed by low-pass filters. The outputs of the smoothing filters are low frequency signals which represent the short-time spectral energy in each of the analyzed bands. These signals represent the slowly varying characteristics of a speech signal. Physically, they are just the average of the energy of the signals from the bandpass filters. The output of the excitation detector (be it conventional pitch detector with voiced-unvoiced decision or a narrow band of the input speech) is transmitted by a suitable frequency or time multiplexing method along with the slowly varying spectrum signals. This information is used by the synthesizer portion of the vocoder to regenerate the input speech. Synthesis is accomplished by amplitude modulating the harmonics of a locally generated source or carrier function by the corresponding low frequency spectrum channels. The harmonics of the excitation function are obtained from a set of filters (not shown in Fig. 1) similar to those used for analyzing the input speech signal. The modulated signals are then band limited in order to remove spurious sidebands introduced by the modulation process. The combined output of these filters is a replica of the input speech. Hence, wideband speech is synthesized from a set of low frequency signals having a relatively narrow overall bandwidth.

This, then, is the complex communication system for which a suitable sampled-data representation is sought.

SAMPLED DATA REPRESENTATION OF A CONTINUOUS SYSTEM

In order that a suitable sampled-data representation be found for the vocoder shown in Fig. 1, it is necessary that the transfer functions be known for the various blocks in the system. For example, the various bandpass and low-pass filters are first represented by their transfer functions in the form:

$$H(s) = \frac{C_o \prod_{k=1}^{K} (s-\alpha_k)}{\prod_{\ell=1}^{L} (s-\beta_\ell)} \qquad (1)$$

where:

C_o = constant

α_ℓ = complex zeros of $H(s)$

β_ℓ = complex poles of $H(s)$

K = degree of numerator

L = degree of denominator

s = complex Laplace transform variable

Conversion of this type of function to an appropriate sampled-data form was accomplished by applying either the z-transform or bilinear z-transformation[5,6] to the above transfer function.

The resulting z-transfer function was then written as a partial fraction expansion in the form,

$$\mathcal{H}(z) = \sum_{m=1}^{M} \frac{A_{1m} z^{-1} + A_{Om}}{B_{2m} z^{-2} + B_{1m} z^{-1} + 1} \qquad (2)$$

where:

$A_{1m}, A_{0m}, B_{2m}, B_{1m}$ are constants

$z^{-1} = \exp(-sT) =$ the unit delay operator

$T =$ unit sampling interval

Each term of the above expansion is then represented in block diagram form as shown in Fig. 2a. Figure 2b shows the realization of the complete transfer function. This technique was used to realize all of the linear transfer functions in the system. (Incidentally, the entire design of such functions is readily carried out by digital computer. Output from such a design program will be presented later.)

The nonlinear function performed by the rectifier is simply,

$$\mathfrak{R}_{out}(z) = |\mathfrak{R}_{in}(z)| \tag{3}$$

while that of the "chopper-modulators" is,

$$\mathfrak{M}_{out}(z) = \mathfrak{M}_{in_1}(z) * SGN(\mathfrak{M}_{in_2}(z)) \tag{4}$$

Similarly appropriate sampled-data representations were made for the excitation detector and generator. This, then, completes the first prerequisite for computer simulation. A BLODIB computer program was prepared next using the above information.

THE BLODIB PROGRAMMING LANGUAGE

As stated earlier, the BLODIB programming language is system oriented in that it closely resembles an engineering block diagram. More specifically, it is a verbal description of the various blocks in the diagram along with parameter values (where required) and information concerning how the blocks are interconnected. In addition, BLODIB makes possible the "construction" or definition of new types of blocks (called MACROS or SUPERS) from the basic types available. (Special boxes not in the basic set of 30 may also be introduced by supplying an external algorithm coded as a FAP subroutine or as a FORTRAN function. However this technique will not be discussed here.)

The use of the MACRO facility as well as the simplicity of BLODIB coding can be demonstrated easily with respect to the realization of the filter transfer functions. Figure 3 shows the actual BLODIB configuration for the basic terms required in the partial fraction expansion given by Eq. (2). Since this "term" or function is required many times, it will be defined as a new type - ADB - and thereafter used as a basic building block. Coding in BLODIB requires giving each block a name (chosen by the programmer); specifying the type of block from the list of basic types, followed by any parameters associated with the block; and finally, specifying to what blocks the output should be connected. Thus the actual coding for Fig. 3 is given in Table I. The first line of coding defines ADB as a SUPER which may have up to 4 inputs (I1, I2 etc.) and has 4 parameters A1, A0, B2, B1. The boxes MIP and MØT are required input/output boxes for SUPERS. The term END obviously designates the end to the coding for ADB. The order of appearance of the boxes within the SUPER is immaterial since the compiler internally orders the final connection matrix in order to produce an efficient program.

The partial fraction expansion terms in the various filter transfer functions were then realized using this definition for ADB. Since all of the filters, bandpass and low-pass, are used in several places, these too were coded as SUPERS. Hence wherever a filter was required, it could be referenced as a single block. In this way, a complete set of filter blocks was built up quickly from the basic BLODIB language.

Figures 4a and 4b show respectively the frequency and time response of one of the bandpass filters required in the simulation. (These graphs along with the BLODIB programming required for simulation were obtained by the computer program mentioned earlier.) The block diagram for this filter is shown in Fig. 5 while Table II presents the actual BLODIB coding.

Having determined the representation for the various filters, the next step required was that of coding each of the 15 channels in the vocoder. From Fig. 1, it is seen that each channel consists of a bandpass filter, a full wave rectifier, a low-pass filter, a "chopper" modulator, and another bandpass filter. In addition, a third bandpass filter is used between the output of the excitation generator and the modulator. A block diagram for one of these channels is shown in Fig. 6 along with the BLODIB coding necessary for simulation. The power of BLODIB with its "SUPER" capability is now readily apparent. Not only is the actual coding for each channel simplified, but any or all of the filters may be changed (by substitution of different SUPERS) without the necessity for recoding the entire system.

Completion of the above second prerequisite is usually all that is required in the programming of an entire system. However, two other features of BLODIB which were also used advantageously will be discussed briefly. These are the modular programming feature called SSUBR and the subroutine feature called SUBR. The SSUBR feature allows a block of BLODIB coding to be used as an external "module" to a main BLODIB program. That is, unlike SUPERs which must be included in the main program, SSUBRs are coded, compiled and loaded separately from the main program.

For example, the MØD (chopper) box shown in Fig. 6 was coded as an SSUBR and used in conjunction with the main vocoder program. This allowed different modulator configurations to be tested by simply "plugging" them into program deck at run time. Multiplexing systems also were tested by using different external SSUBRs.

The SUBR feature permits coding an entire simulation so that it can be controlled by a main or "executive" program. (Such a program may be coded in FORTRAN or FAP.) This permits changing system parameters or using intermediate analysis programs as an integral part of the simulation.

At present, the simulated vocoder consists of a main Fortran program which supplies external parameters and provides the calling sequences for the following BLODIB subroutines (SUBRs): an input routine which allows either conventional or voice-excited vocoder mode to be used; the main vocoder program; and, the output routine. These subroutines in turn, each use BLODIB SSUBRs for the baseband filter, the spectrum spreading network, and transmission coding quantizers. Figure 7 shows how the various programs are interconnected. Variation of parameter or network configuration is accomplished by changing either parameters within a given block or changing the block completely. In this manner, optimization of system performance was achieved by optimizing the appropriate "modules" in the simulation. Hence, the programming of a complete vocoder system is achieved in an extraordinarily short amount of time using BLODIB.

Having completed the first two important prerequisites for simulation - namely sampled-data representation of the vocoder and its translation into the BLODIB programming language - there remains only the final step of preparing input data to be processed by the computer. Preparation of input data for the computer is made by an analog-digital translator as indicated in Fig. 8. The translator, which can sample prerecorded input speech at rates up to 20 kcps, quantizes the input speech signal to one of 4096 levels (including sign). The sampled-and-quantized data is recorded at 800 bits/inch on standard digital magnetic tape in a format used by the BLODIB input/output routines. Running time on the IBM 7094^{II} for typical vocoder simulations is about 100 times real time (i.e., 1 sec of speech requires 100 sec for processing). The results from the simulation programs are similarly recorded on digital tape and then played back through the data translator. Of course, digital input tapes can be used over and over again thereby removing this third prerequisite from succeeding simulations.

Thus far, BLODIB computer simulations have indicated: what filter designs yield good synthesized speech; what type of quantizing and sampling rates might be satisfactorily used on the spectrum channel signals, how many spectrum channels are required, etc. As a result of these findings, improvement of system performance has been accomplished in a relatively short time and without the inherent delay and cost of building complex equipment.

SUMMARY

Digital computer simulation of communication systems has been simplified and made more flexible by use of the BLODIB programming language.

Three prerequisites to actual simulation have been presented. They are:

1) The finding of an appropriate sampled-data representation for the particular communication system,

2) The preparation of a BLODIB computer program for the above sampled-data representation and,

3) The preparation of input data for processing by the program.

Details with respect to the BLODIB programming of the second step above were illustrated by their application in the simulation of a vocoder system. Hence new and complex communication systems can be evaluated, optimized and redesigned quickly and efficiently before final construction in hardware.

REFERENCES

1. Karafin, B. J., The New Block Diagram Compiler for Simulation of Sampled-Data Systems, AFIPS 1965 FJCC.
2. Kelly, J. L., Jr., Lochbaum, C., Vyssotsky, V. A., "A Block Diagram Compiler, B.S.T.J., 40, May, 1961, pp. 669-676.
3. Schroeder, M. R., Logan, B. F., Prestigiacomo, A. J., "New Methods for Speech Analysis-Synthesis and Bandwidth Compression," Fourth International Congress on Acoustics, Copenhagen, August 21-28, 1962.
4. Golden, R. M., "Digital Computer Simulation of a Sampled-Data Voice-Excited Vocoder," Journal of the Acoustical Society of America, Vol. 35, 9 pp. 1358-1366, Sept. 1963.
5. Kaiser, J. F., "Design Methods for Sampled-Data Filters," Proc. First Allerten Conference on Circuit and System Theory, Nov. 1963, Monticello, Illinois.
6. Golden, R. M., Kaiser, J. F., "Design of Wideband Sampled-Data Filters," The Bell System Technical Journal, Vol. XLIII, 4, July, 1964, Pt. 2, pp. 1533-1546.

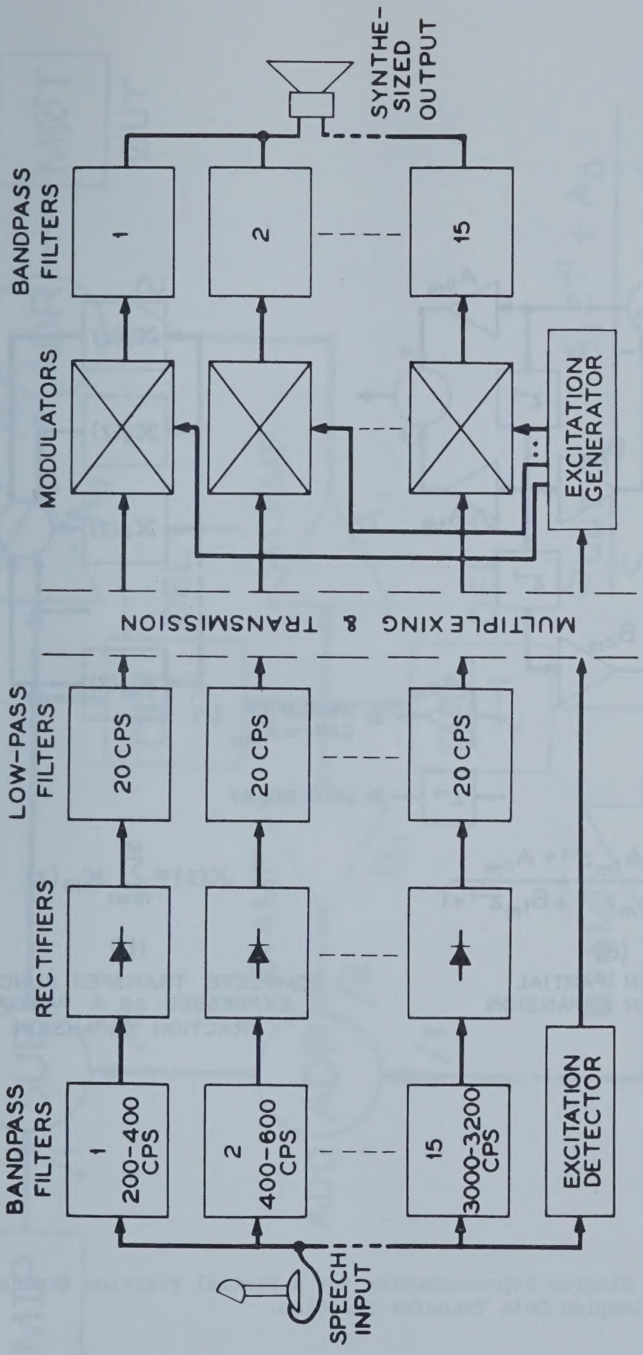
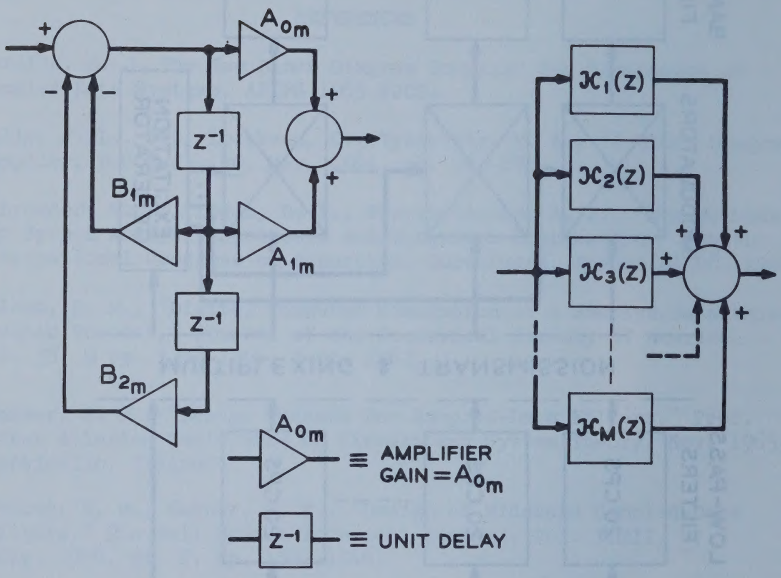

SPEECH INPUT
BANDPASS FILTERS
1 200-400 CPS
2 400-600 CPS
15 3000-3200 CPS
EXCITATION DETECTOR
RECTIFIERS
LOW-PASS FILTERS
20 CPS
20 CPS
20 CPS
MULTIPLEXING & TRANSMISSION
MODULATORS
EXCITATION GENERATOR
BANDPASS FILTERS
1
2
15
SYNTHESIZED OUTPUT

Fig. 1 Functional Block Diagram of Vocoder.


+
−
−
A_{0m}
z^{-1}
+
+
B_{1m}
A_{1m}
z^{-1}
B_{2m}
A_{0m}
≡ AMPLIFIER GAIN = A_{0m}
z^{-1}
≡ UNIT DELAY
$\mathcal{H}_1(z)$
$\mathcal{H}_2(z)$
$\mathcal{H}_3(z)$
$\mathcal{H}_M(z)$
+
+
+
+

$$\mathcal{H}_m(z) = \frac{A_{1m} z^{-1} + A_{0m}}{B_{2m} z^{-2} + B_{1m} z^{-1} + 1}$$

(a)
TERM IN PARTIAL FRACTION EXPANSION

$$\mathcal{H}(z) = \sum_{m=1}^{M} \mathcal{H}_m(z)$$

(b)
COMPLETE TRANSFER FUNCTION EXPRESSED AS A PARTIAL FRACTION EXPANSION

Fig. 2 Block Diagram Representation for a Partial Fraction Expansion of a Sampled-Data Transfer Function.

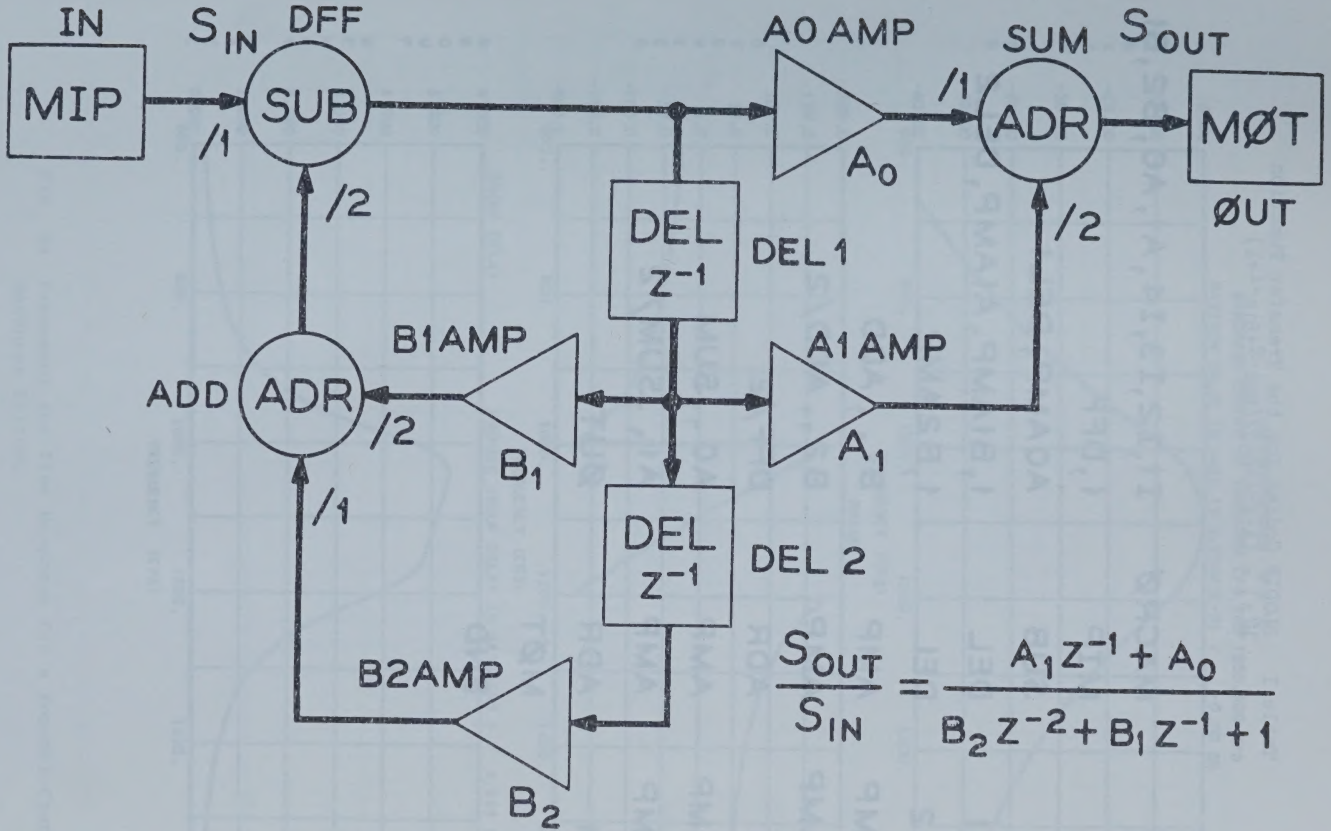

IN
MIP
S_{IN}
DFF
SUB
/1
/2
A0 AMP
A_0
SUM
S_{OUT}
/1
ADR
/2
MØT
ØUT
DEL
z^{-1}
DEL 1
B1AMP
B_1
ADD
ADR
/2
/1
A1 AMP
A_1
DEL
z^{-1}
DEL 2
B2AMP
B_2
$\frac{S_{OUT}}{S_{IN}} = \frac{A_1 Z^{-1} + A_0}{B_2 Z^{-2} + B_1 Z^{-1} + 1}$

Fig. 3 BLODIB Representation for the Transfer Function $\mathfrak{H}(z) = (A_1z^{-1}+A_0)/(B_2z^{-2}+B_1z^{-1}+1)$.

Table I BLODIB Coding for the Transfer Function $\mathcal{H}(z) = (A_1 z^{-1}+A_0)/(B_2 z^{-2}+B_1 z^{-1}+1)$.

```
ADB      MACRØ   I1,I2,I3,I4,A1,A0,B2,B1
IN       MIP     1,DFF
DFF      SUB     A0AMP,DEL1
DEL1     DEL     1,B1AMP,A1AMP,DEL2
DEL2     DEL     1,B2AMP
B1AMP    AMP     B1,,ADD
B2AMP    AMP     B2,,ADD/2
ADD      ADR     DFF/2
A0AMP    AMP     A0,,SUM
A1AMP    AMP     A1,,SUM/2
SUM      ADR     ØUT
ØUT      MØT
         END
```

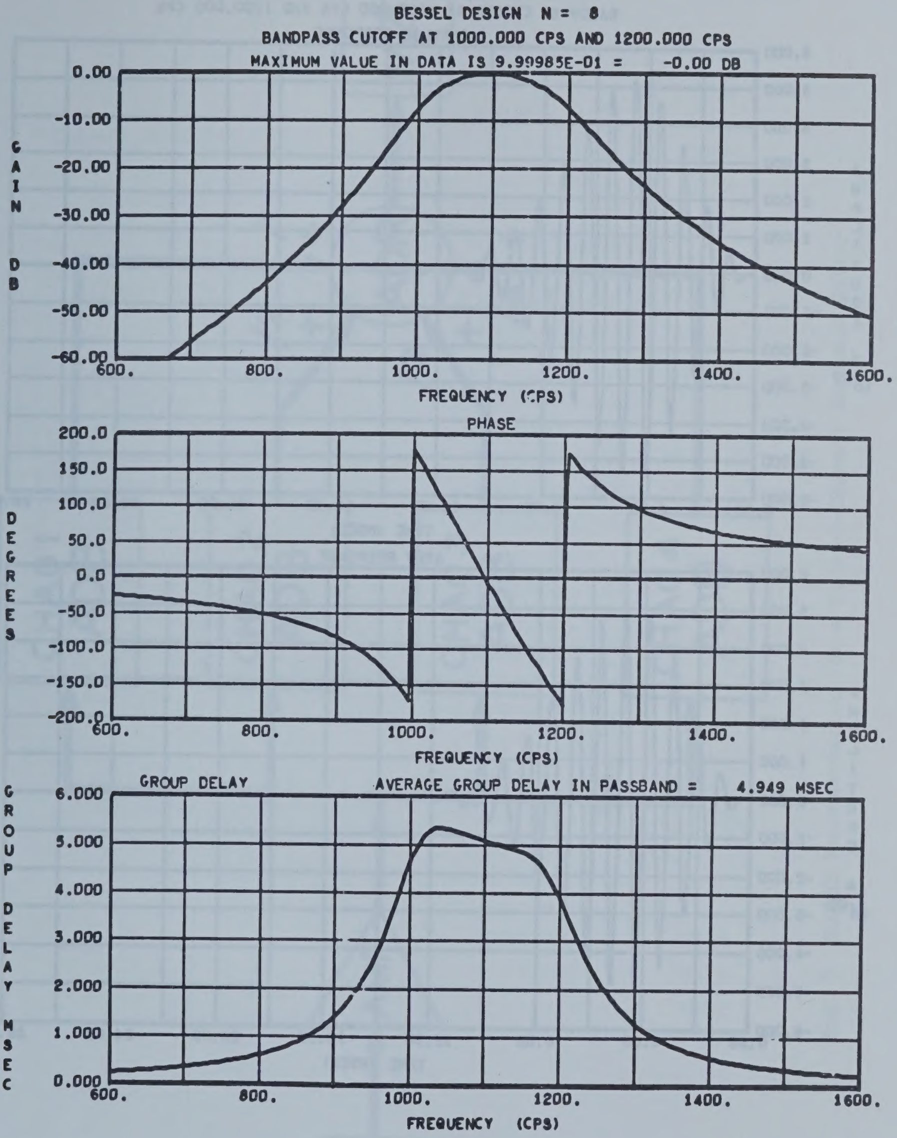
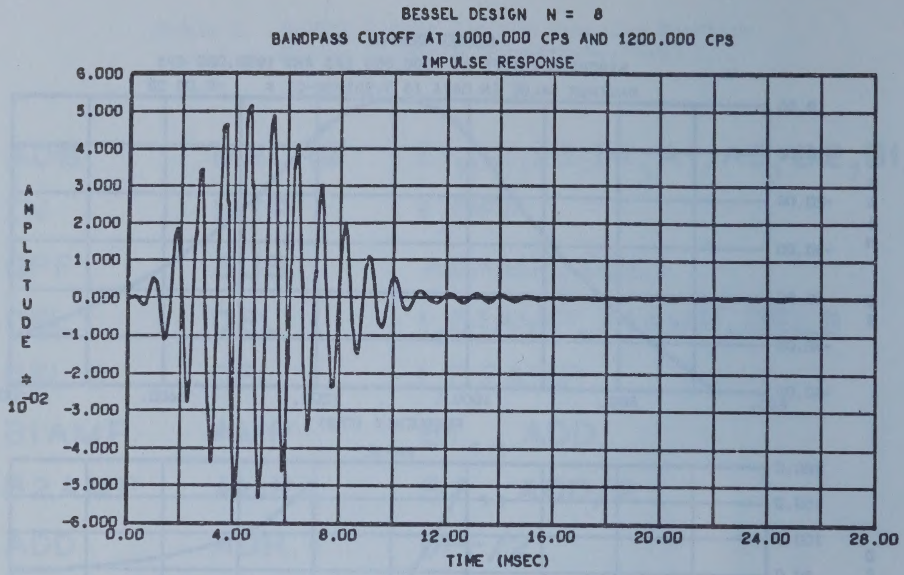


Fig. 4a Frequency and Time Response for a Vocoder-Channel Bandpass Filter.



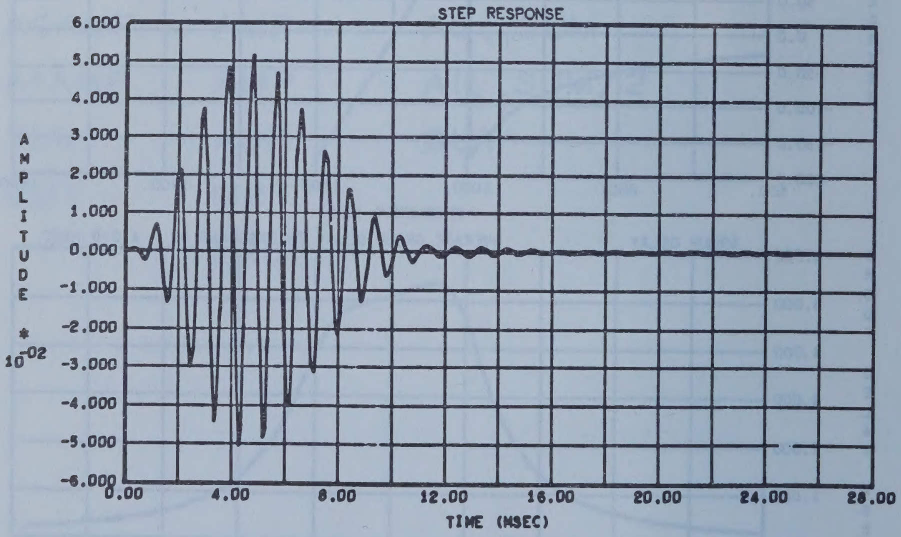


Fig. 4b Frequency and Time Response for a Vocoder-Channel Bandpass Filter.

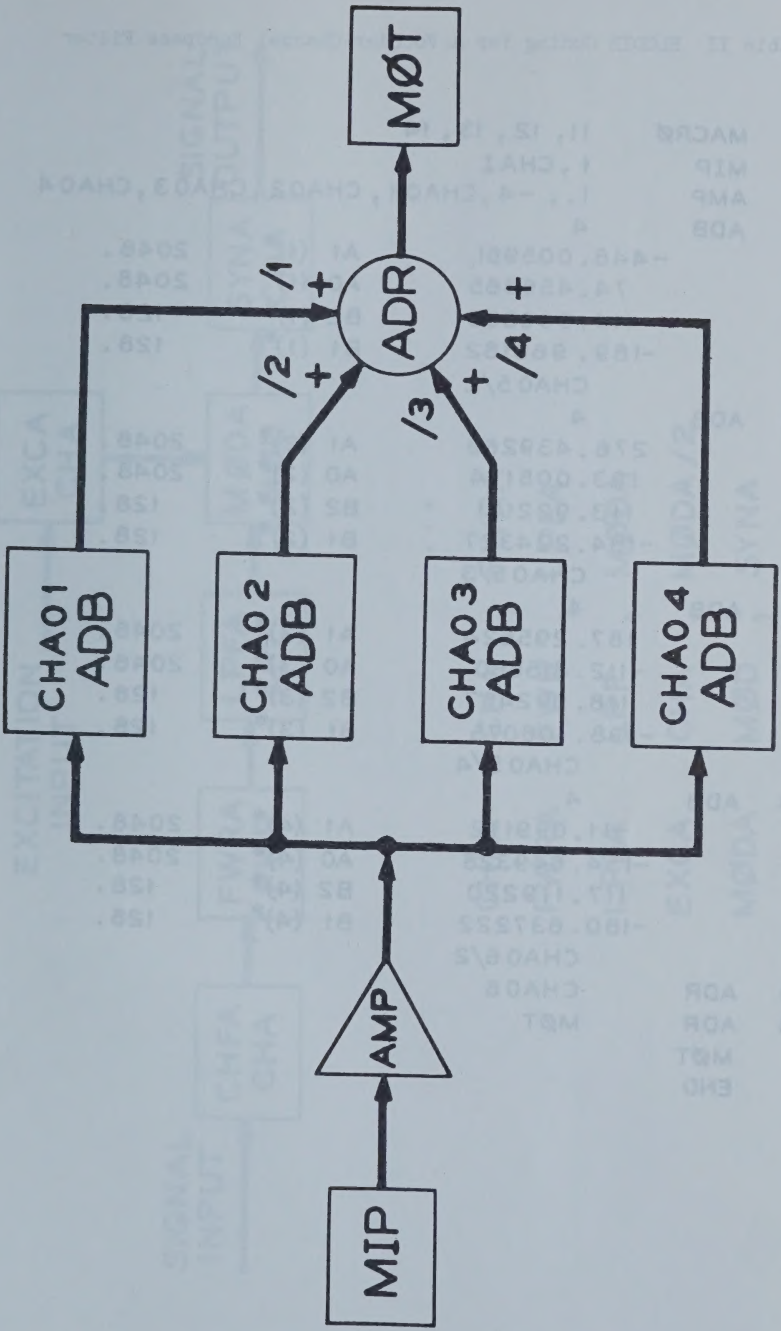
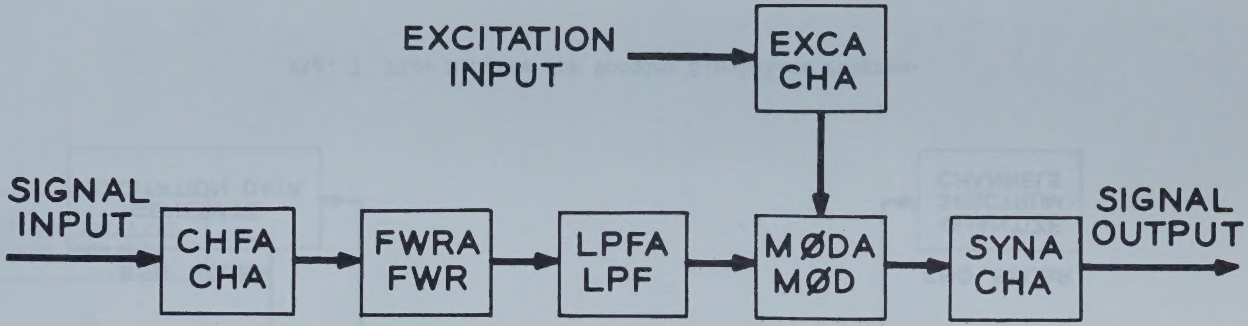

MIP
AMP
CHA01 ADB
CHA02 ADB
CHA03 ADB
CHA04 ADB
/1 +
/2 +
/3 +
/4 +
ADR
MØT

Fig. 5 BLODIB Representation for a Vocoder-Channel Bandpass Filter.

Table II BLODIB Coding for a Vocoder-Channel Bandpass Filter

```
CHA       MACRØ     11, 12, 13, 14
IN        MIP       1, CHAI
CHAI      AMP       1., -4, CHA01, CHA02, CHA03, CHA04
CHA01     ADB       4
                 -448.005981        A1  (1)*      2048.
                   74.456865        A0  (1)*      2048.
                  114.566368        B2  (1)*       128.
                 -189.968182        B1  (1)*       128.
                    CHA05/2
CHA02     ADB       4
                  276.439266        A1  (2)*      2048.
                  193.008114        A0  (2)*      2048.
                  113.922181        B2  (2)*       128.
                 -184.224327        B1  (2)*       128.
                    CHA05/3
CHA03     ADB       4
                  187.295624        A1  (3)*      2048.
                 -112.815660        A0  (3)*      2048.
                  118.592467        B2  (3)*       128.
                 -198.106075        B1  (3)*       128.
                    CHA05/4
CHA04     ADB       4
                  -11.099112        A1  (4)*      2048.
                 -154.649326        A0  (4)*      2048.
                  117.119220        B2  (4)*       128.
                 -180.637222        B1  (4)*       128.
                    CHA06/2
CHA05     ADR       CHA06
CHA06     ADR       MØT
MØT       MØT
          END
```


EXCITATION INPUT
EXCA
CHA
SIGNAL INPUT
CHFA
CHA
FWRA
FWR
LPFA
LPF
M0DA
M0D
SYNA
CHA
SIGNAL OUTPUT

```
                        . . .
CHFA     CHA    ,  FWRA
FWRA     FWR    ,  LPFA
LPFA     LPF    ,  M0DA
EXCA     CHA    ,  M0DA/2
M0DA     M0D    ,  SYNA
SYNA     CHA    ,   . . .
```

Fig. 6 BLODIB Representation and Coding for a Complete Vocoder Channel.

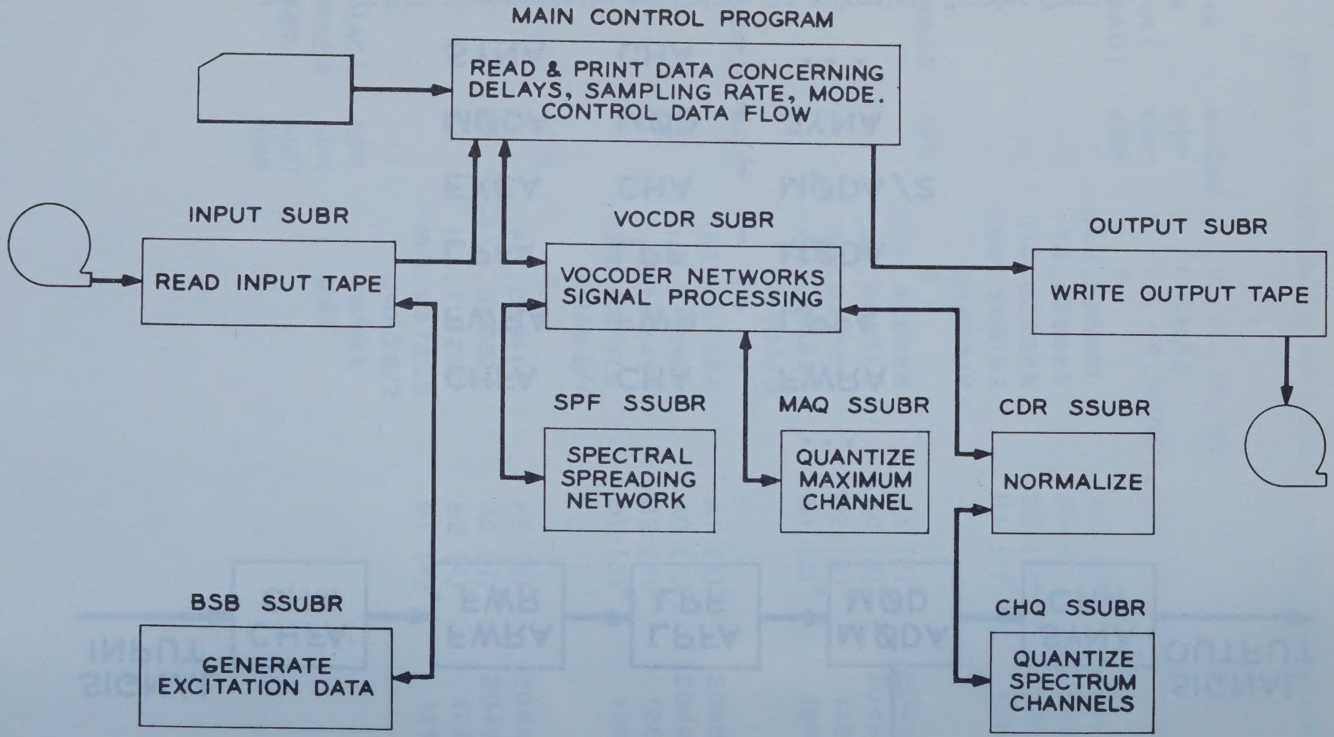

MAIN CONTROL PROGRAM
READ & PRINT DATA CONCERNING DELAYS, SAMPLING RATE, MODE. CONTROL DATA FLOW
INPUT SUBR
READ INPUT TAPE
VOCDR SUBR
VOCODER NETWORKS SIGNAL PROCESSING
OUTPUT SUBR
WRITE OUTPUT TAPE
SPF SSUBR
SPECTRAL SPREADING NETWORK
MAQ SSUBR
QUANTIZE MAXIMUM CHANNEL
CDR SSUBR
NORMALIZE
BSB SSUBR
GENERATE EXCITATION DATA
CHQ SSUBR
QUANTIZE SPECTRUM CHANNELS

Fig. 7 Flow Diagram for Vocoder Simulation Program.

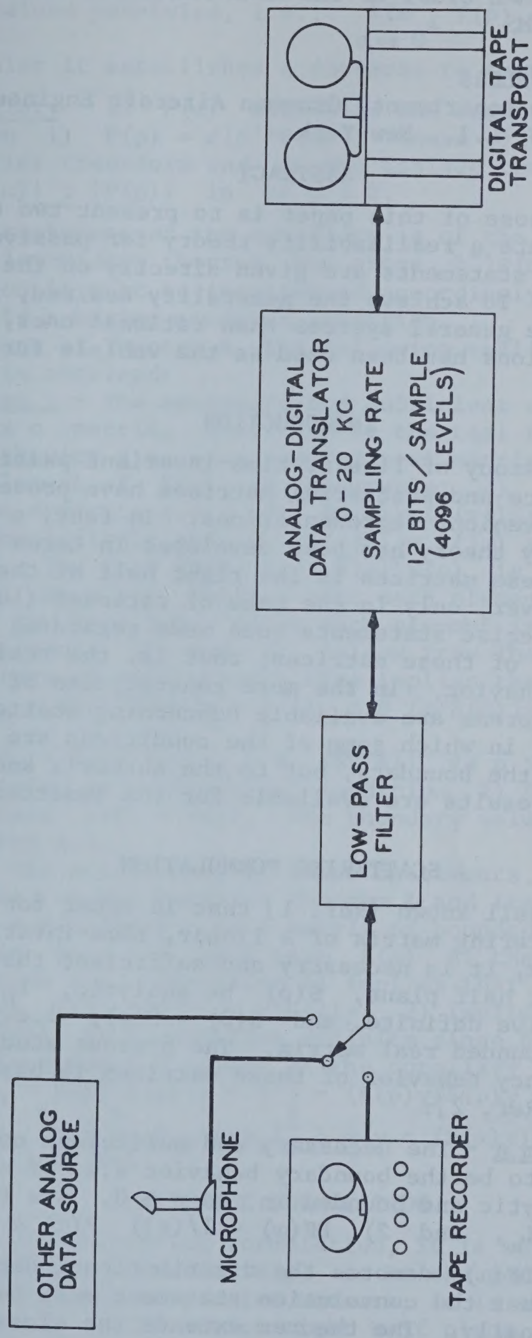

OTHER ANALOG DATA SOURCE
MICROPHONE
TAPE RECORDER
LOW-PASS FILTER
ANALOG-DIGITAL DATA TRANSLATOR
0-20 KC
SAMPLING RATE
12 BITS/SAMPLE
(4096 LEVELS)
DIGITAL TAPE TRANSPORT

Fig. 8 Preparation of Digital Input Data for Use with Computer Simulation of Speech Communication Systems.

A DISTRIBUTIONAL STUDY OF THE REAL FREQUENCY BEHAVIOR OF PASSIVE SYSTEMS

M. R. WOHLERS
Research Department, Grumman Aircraft Engineering Corp.
Bethpage, L.I., New York

ABSTRACT

The purpose of this paper is to present two theorems that constitute a realizability theory for passive n-ports in which all statements are given directly on the real frequency axis. To achieve the generality desired, i.e., to consider more general systems than rational ones, the theory of distributions has been used as the vehicle for the analysis.

INTRODUCTION

In the study of linear, time-invariant passive systems, the immittance and scattering matrices have proven to be the most convenient representations. In fact, a complete realizability theory has been developed in terms of the behavior of these matrices in the right half of the complex plane. However, only in the case of rational (lumped) systems have precise statements been made regarding the boundary behavior of these matrices; that is, the real frequency ($p = j\omega$) behavior. In the more general case of irrational systems, theorems are available concerning scattering matrices [Ref. 1] in which some of the conditions are stated directly on the boundary, but to the author's knowledge, no comparable results are available for the immittance description.

SCATTERING FORMULATION

It is well known [Ref. 1] that in order for $S(p)$ to be the scattering matrix of a linear, time-invariant, passive network, it is necessary and sufficient that in the strict right half plane, $S(p)$ be analytic, $1_n - \bar{S}(p)^T S(p)$ be nonnegative definite, and $S(\bar{p}) = \bar{S}(p)$, i.e., $S(p)$ must be a bounded real matrix. The present study of the real frequency behavior of these matrices is based on the following [Ref. 2]:

Theorem A - The necessary and sufficient conditions for $F(\omega)$ to be the boundary behavior a.e. of a function, $F(p)$, analytic and bounded in $\operatorname{Re} p > 0$, are that: 1) $F(\omega) \in L_\infty$, and 2) $DF(\omega) = 1/(\pi j)\ F(\omega) * D\ \mathrm{pv}\ 1/\omega$.

Note that $DF(\omega)$ denotes the distributional derivative of $F(\omega)$ and that the convolution statement must be interpreted distributionally. The theorem extends the classic result of

Fatou which states that bounded analytic functions have boundary values pointwise, i.e., $\lim_{\sigma \to 0^+} F(p) = F(\omega)$ a.e., in particular it establishes a converse to that result.

Corollary - If $F(\omega)$ satisfies the condition of Theorem A, then 1) $F(p) = \mathscr{L}[\mathfrak{F}^{-1}F(\omega)]$, where $\mathfrak{F}^{-1}$ is the inverse Fourier transform and $\mathscr{L}$ the Laplace transform, and 2) $\max|F(\omega)| \geq |F(p)|$ in $\operatorname{Re} p > 0$.

The first statement of the corollary is of the form of a classic Paley-Wiener theorem, but since $\mathfrak{F}^{-1}F(\omega)$ may be a distribution it must be interpreted accordingly. The second statement is a maximum-modulus condition.

By means of Theorem A, the following realizability statement is obtained:

Theorem 1 - The necessary and sufficient conditions for an $n \times n$ matrix, $S(\omega)$, to be the real frequency boundary behavior a.e. of a bounded real matrix, are that: 1) each element of $S(\omega)$ be measurable; 2) $1 - \bar{S}^{T}S$ be nonnegative definite a.e.; 3) $DS(\omega) = 1/(\pi j)S(\omega) * D \text{ pv } 1/\omega$ (scalar convolution); and 4) $S(-j\omega) = \bar{S}(j\omega)$ a.e.

Proof - The fact that $1_n - \bar{S}(p)^{T}S(p)$ is nonnegative definite in $\operatorname{Re} p > 0$ implies that each element of $S(p)$ is bounded in $\operatorname{Re} p > 0$. Since each element is also analytic, statements 1) and 3) follow from Theorem A. The bounded nature of each element also implies that

$$\lim_{\sigma \to 0} \bar{\underset{\sim}{y}}^{T}[1_n - \bar{S}(p)^{T}S(p)]\underset{\sim}{y} = \bar{\underset{\sim}{y}}^{T}[1_n - \bar{S}(\omega)^{T}S(\omega)]\underset{\sim}{y} \text{ a.e., and}$$

since $\bar{y}^{T}[1_n - \bar{S}(p)^{T}S(p)]y \geq 0$ for all $\operatorname{Re} p > 0$, the limit must also be nonnegative; condition 2) follows. Finally, since $S(\bar{p}) = \bar{S}(p)$, the boundary values have the same property a.e.

As to the sufficiency of these statements, 1), 2), and 3) guarantee, on the basis of Theorem A and its corollary, that $S(p) = \mathscr{L}[\mathfrak{F}^{-1}S(\omega)]$ is a matrix of bounded analytic functions in $\operatorname{Re} p > 0$ that have $S(\omega)$ as their pointwise boundary value a.e.. Condition 4) implies that $\mathfrak{F}^{-1}S(\omega)$ is real so that $S(\bar{p}) = \bar{S}(p)$. Finally, 2) implies that $1_n - \bar{S}(p)^{T}S(p)$ is nonnegative definite since by appealing to the maximum modulus aspect of the corollary we conclude that $\bar{\underset{\sim}{y}}^{T}[1_n - \bar{S}(p)^{T}S(p)]\underset{\sim}{y} = \bar{\underset{\sim}{y}}^{T}\underset{\sim}{y} - (\overline{S(p)\underset{\sim}{y}})^{T}S(p)\underset{\sim}{y} =$

$$\sum_{k=1}^{n} |y_k|^2 - \sum_{j=1}^{n} \left|\sum_{k=1}^{n} s_{jk}(p)y_k\right|^2 \geq \bar{\underset{\sim}{y}}^{T}\underset{\sim}{y} - (\overline{S(\omega)\underset{\sim}{y}})^{T}S(\omega)\underset{\sim}{y} \geq 0.$$

QED

IMMITTANCE FORMULATION

As in the scattering formulation, it is well known [Ref. 3] that for a matrix $W(p)$ to be the immittance of a linear, time-invariant passive network, it is necessary and sufficient that in the strict right half plane, $W(p)$ be analytic, $\bar{W}(p)^T + W(p)$ be nonnegative definite, and $W(\bar{p}) = \bar{W}(p)$, i.e., $W(p)$ must be a positive real matrix. The study of the boundary behavior of positive real matrices is based on Youla's generalization of Cauer's representation of positive real functions [Refs. 3 and 4]:

Theorem B - The necessary and sufficient condition that an $n \times n$ matrix $W(p)$ be positive real is that for $\mathrm{Re}\, p > 0$,

$$W(p) = Q + Ap + \int_{-\infty}^{\infty} \frac{1 - pj\tau}{p - j\tau}\, dK(\tau) \quad ,$$

where Q is a real, constant, skew-symmetric $n \times n$ matrix; A is a real, constant, symmetric, nonnegative definite matrix; and $K(\tau)$ is an $n \times n$ matrix, each of whose elements are of bounded variation on the entire line $-\infty < \tau < \infty$, satisfying $\bar{K}^T = K$, $-\bar{K}(-\tau) = K$, and with $\bar{\underset{\sim}{a}}^T K(\tau)\underset{\sim}{a}$ being a real, bounded nondecreasing function of τ for any complex constant $n \times 1$ vector $\underset{\sim}{a}$.

When we attempt to study the real frequency behavior of the positive real matrix, using classical techniques, we run into immediate difficulties. For example, the simple positive real function $1/p$ has a pointwise boundary value of $1/(j\omega)$ a.e. and so $\lim\limits_{\sigma \to 0} \mathrm{Re}\, 1/p = 0$ a.e. The inverse Fourier transform of $1/(j\omega)$ is $\frac{1}{2}\,\mathrm{sgn}\, t$ which happens to equal $\frac{1}{2}$ for $t > 0$ and we are forced into the position (on the basis of using the pointwise boundary value) that the impulse response of the network is not causal. However, it is well known that $1/p = \mathcal{L}[u(t)]$ and that, distributionally, $\mathfrak{F}[u(t)] = \pi\delta + \mathrm{pv}\ 1/(j\omega) = 1/(j\omega)$ a.e. The resolution of this difficulty is to consider the distributional or weak boundary value of the function $1/p$ which is $\pi\delta + \mathrm{pv}\ 1/(j\omega)$. A complete study of distributional boundary values may be made on the basis of the following generalization of the Paley-Wiener theorem [Ref. 5]:

Theorem C - The Laplace transform of a distribution in S_+' (the dual space of S_+) converges in the S' topology to the Fourier transform of the same distribution, i.e.,

$$\lim_{\sigma \to 0} < F(\sigma + j\omega),\ \varphi(\omega) > = < F_\omega,\ \varphi(\omega) >$$

for all $\varphi(\omega) \in S$.

Note that $\varphi \in S$ if it is infinitely differentiable and it, and all its derivatives, vanish faster than any polynomial at infinity, whereas $\varphi \in S_+$ if $\varphi \in S$ and $\varphi = 0,\ t < 0$.

In the particular case when $F(p)$ is positive real, we will establish that it has boundary values in the S' topology which completely characterize it. In fact, Theorems B and C form the basis of the following realizability statement:

Theorem 2 - The necessary and sufficient condition that an $n \times n$ matrix of distributions W_ω be the boundary behavior (in the S' topology) of a positive real matrix is that

$$W_\omega = Q + j\omega\left[A + \int_{-\infty}^{\infty} dK\right] + \pi(1+\omega^2)DK + (1+\omega^2)\left[DK * \text{pv}\,\frac{1}{j\omega}\right].$$

where Q, A, and K are defined in Theorem C and in particular $\bar{W}^T + W = 2\pi(1+\omega^2)DK$.

Proof - The necessity will be established if we demonstrate that W_ω given above is the S' boundary value of the representation for $W(p)$ given in Theorem C. The first two terms of that representation converge in their given form as $p \to j\omega$. The third term may be rewritten, by noting that $(1 - pj\tau)/(p - j\tau) = p + (1 - p^2)/(p - j\tau)$, in the form $p \int_{-\infty}^{\infty} dK + (1 - p^2) \int_{-\infty}^{\infty} 1/(p - j\tau)\,dK$, where we have observed that the bounded variation of $K(\tau)$ implies that $\int_{-\infty}^{\infty} dK$ exists. The third term of the expression for W_ω is then obtained directly. Finally, consider for all $\varphi \in S$, $\lim_{\sigma \to 0} < (1 - p^2) \int_{-\infty}^{\infty} 1/(p - j\tau)\,dK, \varphi(\omega) > =$ $\lim_{\sigma \to 0} \int_{-\infty}^{\infty} dK \int_{-\infty}^{\infty} (1 - p^2)/(p - j\tau)\, \varphi(\omega)\,d\omega$ where the classic Fubini theorem justifies the interchange. However, Parseval's relation allows one to write $\int_{-\infty}^{\infty} (1-p^2)/(p-j\tau)\,\varphi(\omega)\,d\omega =$ $2\pi \int_{-\infty}^{\infty} \hat{\varphi}_\sigma(t) e^{j\tau t} e^{-\sigma t} u(t)\,dt$, where $\hat{\varphi}_\sigma(-t) =$ $\mathfrak{F}^{-1}[(1 - p^2)\varphi(\omega)] \in S$ for every $\sigma \geq 0$, and $\mathfrak{F}^{-1}[1/(p-j\tau)] =$ $e^{j\tau t} e^{-\sigma t} u(t)$. Therefore, $|\int_{-\infty}^{\infty} (1 - p^2)/(p - j\tau)\varphi(\omega)\,d\omega| =$ $2\pi |\int_{-\infty}^{\infty} \hat{\varphi}_\sigma(t) e^{j\tau t} e^{-\sigma t} u(t)\,dt| \leq 2\pi \int_0^{\infty} |\hat{\varphi}_\sigma| e^{-\sigma t}\,dt \leq M$ for all $\sigma \geq 0$. Then, the dominated convergence theorem allows one to conclude that

$\lim_{\sigma \to 0} \int_{-\infty}^{\infty} dK \int_{-\infty}^{\infty} (1 - p^2)/(p - j\tau)\, \varphi(\omega)\,d\omega =$

$$\int_{-\infty}^{\infty} dK \lim_{\sigma \to 0} \int_{-\infty}^{\infty} (1 - p^2)/(p - j\tau)\,\varphi(\omega)\,d\omega =$$

$$\int_{-\infty}^{\infty} dK \left[pv \int_{-\infty}^{\infty} [1 + (\omega + \tau)^2]/(j\omega)\,\varphi(\omega + \tau)\,d\omega + (1 + \tau^2)\pi\varphi(\tau) \right].$$

Thus, we have shown that

$$\lim_{\sigma \to 0} < (1 - p^2) \int_{-\infty}^{\infty} 1/(p - j\tau)\,dK, \varphi(\omega) > =$$

$$\int_{-\infty}^{\infty} (1 + \tau^2)\pi\varphi(\tau)\,dK +$$

$$\int_{-\infty}^{\infty} dK\; pv \int_{-\infty}^{\infty} [1 + (\omega + \tau)^2]/(j\omega)\;\varphi(\omega + \tau)\,d\tau =$$

$$< \pi(1 + \omega^2)DK(\omega), \varphi(\omega) > + < (1 + \omega^2)\left(DK * pv\; 1/(j\omega) \right), \varphi(\omega) >.$$

The sufficiency of the theorem follows from the remark that $\overline{W}^{T} + W = 2\pi(1 + \omega^2)DK$ (which may be established directly from the expression for W_{ω}) because this allows one to identify $K(\omega)$, given W_{ω}, and then by inspection to determine Q, and A from W_{ω}. The conditions of the theorem guarantee that the Q, A, and K so identified will satisfy the requirements of Theorem B and thus the W(p) formed from them, by means of Youla's representation, will be positive real. The above discussion proves that the W(p) so formed will converge in the S' topology to the originally given distribution. QED

CONCLUSION

Theorems 1 and 2 constitute two real-frequency characterizations of n-port passive systems which may be of interest, because of their conciseness, to those involved in the study of such systems. The theorems are extensions to the n-port case of results obtained for one-ports and described in Ref. 6.

ACKNOWLEDGMENT

This work was supported in part by the Air Force Office of Scientific Research of the Office of Aerospace Research under Contract AF 49(638)-1512.

REFERENCES

1. D. C. Youla, L. J. Castriota, and H. J. Carlin, "Bounded Real Scattering Matrices and the Foundations of Linear Passive Network Theory," IRE Trans. on Circuit Theory, Vol. CT-6, March 1959.

2. E. J. Beltrami, and M. R. Wohlers, "Distributional Boundary Values of Functions Holomorphic in a Half Plane," to appear in the Journal of Math. and Mechanics.

3. A. Zemanian, "An N-port Realizability Theory Based on the Theory of Distributions," IEEE Trans. on Circuit Theory, Vol. CT-10, June 1963.

4. D. C. Youla, "Representation Theory of Linear Passive Networks," Microwave Research Inst., Polytechnic Institute of Brooklyn, Report No. R-655-58, May 1958.

5. E. J. Beltrami, and M. R. Wohlers, "Distributional Boundary Value Theorems and Hilbert Transforms," Archive for Rational Mechanics and Analysis, Vol. 18, No. 4, 1965.

6. M. Wohlers, and E. Beltrami, "Distribution Theory as the Basis of Generalized Passive Network Analysis," IEEE Trans. on Circuit Theory, Vol. CT-12, June 1965.

IMMITTANCES OF REDUCED DEGREE AND NON-POSITIVE REAL IMMITTANCES IN NETWORK SYNTHESIS

JOHN A. C. BINGHAM
Stanford Linear Accelerator Center
Stanford University
Stanford, California

ABSTRACT

This paper discusses the implications of the occurrence in network synthesis of reactive immittances of reduced degree -- these immittances were originally mentioned by Darlington but apparently have not been much studied since -- and shows with the aid of examples how such occurrences can lead to apparently paradoxical results. The paradoxes are resolved and some conclusions are drawn about the relative accuracies of methods of ladder realization.

MORE ABOUT GOOD AND BAD TWO ELEMENT KIND STRUCTURES*

HARRY B. LEE
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts

ABSTRACT

The results presented by the author at the 1964 Allerton Conference are extended to cover two element kind structures which may contain either

a) one element kind loops, or
b) one element kind cut sets

but not both.

INTRODUCTION

In a recent paper (1,2) the author discussed natural frequencies in the so-called complementary tree structures. Specifically he formulated a simple test which enables one to determine whether or not any particular complementary tree structure has independent natural frequencies.

It is the purpose of the present paper to extend these earlier results. Here we consider two more general types of structures and establish analogous tests for independence of natural frequencies. The specific structure types considered are as follows.

Type I: Two element kind structures which may contain one element kind loops, but which contain no one element kind cut sets.

Type II: Two element kind structures which may contain one element kind cut sets, but which contain no one element kind loops.

A simple rule is developed which enables one to determine whether or not any particular structure of either type has independent natural frequencies.

In the author's earlier paper, all structures are subject to the following conditions.

i) No one element kind loops are allowed.

ii) No one element kind cut sets are allowed.

Thus the results of the present paper generalize the earlier results to the extent that either one, but not both, of these conditions may be dropped.

As in the previous paper, we base our considerations upon RL structures. The results are fully applicable to RC and LC structures, however.

In addition, we continue to refer to RL structures with independent natural frequencies as good structures, and to RL structures with non-independent natural frequencies as bad structures. That is, we adhere to the following definition.

Definition: If the elements of an RL structure of complexity n can be adjusted to positive values so as to make the structure realize any desired set of natural frequencies $s = -\sigma_1, -\sigma_2, \ldots -\sigma_n$ $(\sigma_n \geq \sigma_{n-1} \cdots \geq \sigma_1 > 0)$ within any specified uniform tolerance $\epsilon > 0$, then the structure is said to be good; otherwise the structure is bad.

TYPE I STRUCTURES

We begin by considering Type I structures. Thus let S_i be any (unexcited) Type I structure of complexity i. The nodal equations for S_i always can be written in the form

$$YE = 0 \tag{1}$$

The column matrix E has for its elements the various independent node voltages. The coefficient matrix Y has for its elements signed sums of the branch admittances of S_i. More specifically

I. The typical diagonal element y_{jj} is given by

$$y_{jj} = g_{jj} + \frac{1}{s}\gamma_{jj} \tag{2}$$

where g_{jj} and γ_{jj} denote, respectively, the total conductance and the total reciprocal inductance touching the j^{th} node of S_i.

II. The typical off-diagonal element y_{jk} is given by

$$y_{jk} = -g_{jk} - \frac{1}{s}\gamma_{jk} \tag{3}$$

where g_{jk} and γ_{jk} denote, respectively, the conductance and reciprocal inductance common to the j^{th} and k^{th} nodes of S_i.

The admittance matrix Y in (1) can be separated into conductive and reciprocal inductive components as follows.

$$Y = G + \frac{1}{s}\Gamma . \tag{4}$$

Because Type I structures contain no cut sets of inductances, a resistive tree always is present. Thus the matrix G is positive definite. Similarly, the fact that Type I structures contain no cut sets of resistance guarantees the presence of an inductive tree. Thus the reciprocal inductance matrix Γ also is positive definite.

The fact that both the G and the Γ matrices in (4) are positive definite enables one to apply many standard matrix theorems to G and Γ. Principal among these is the following theorem (3).

Theorem I: Let A and B be square matrices which are real, symmetric and positive definite. In addition, let X_tAX and X_tBX be the quadratic forms of A and B.

If $\lambda_1, \lambda_2 \ldots \lambda_n$ $(\lambda_n \geq \lambda_{n-1} \ldots \geq \lambda_1 > 0)$ denote the roots of the characteristic equation $|A - \lambda B| = 0$, then the maximum and minimum values of the quotient X_tAX/X_tBX are given respectively by

$$\max_{X\neq 0}. \left[\frac{X_tAX}{X_tBX}\right] = \lambda_n$$

$$\min_{X\neq 0}\left[\frac{X_tAX}{X_tBX}\right] = \lambda_1 \quad .$$

If Theorem I is translated so that it applies to the G and Γ matrices, and if it is recognized that the solutions $\lambda = \lambda_n$ and $\lambda = \lambda_1$ of the characteristic equation $|\Gamma - \lambda G| = 0$ are simply the negatives of the largest and the smallest natural frequencies of S_i, then the following useful result is obtained.

Theorem II: The natural frequencies of any Type I structure which are maximal and minimal in absolute value are given respectively by

$$s_{max} = -\max_{X\neq 0}\left[\frac{X_t\Gamma X}{X_tGX}\right]$$

$$s_{min} = -\min_{X\neq 0}\left[\frac{X_t\Gamma X}{X_tGX}\right] .$$

Theorem II leads more or less directly to the following three theorems.

Theorem III: The elements of the G and Γ matrices satisfy the relationship

$$|s_{max}|(g_{jj} \pm 2g_{jk} + g_{kk}) \geq \gamma_{jj} \pm 2\gamma_{jk} + \gamma_{kk} \geq |s_{min}|(g_{jj} \pm 2g_{jk} + g_{kk}) \qquad (5)$$

where s_{max} and s_{min} refer to the natural frequencies of Theorem II.

Proof: Clearly, the following relationship holds for any non-zero column matrix W.

$$\max_{X}\left[\frac{X_t\Gamma X}{X_tGX}\right] \geq \frac{W_t\Gamma W}{W_tGW} \geq \min_{X}\left[\frac{X_t\Gamma X}{X_tGX}\right] . \qquad (6)$$

For a W having $w_j = 1$, $w_k = \mp 1$, and all other elements equal to zero, (6) becomes

$$\max_{X}\left[\frac{X_t\Gamma X}{X_tGX}\right] \geq \frac{\gamma_{jj} \pm 2\gamma_{jk} + \gamma_{kk}}{g_{jj} \pm 2g_{jk} + g_{kk}} \geq \min_{X}\left[\frac{X_t\Gamma X}{X_tGX}\right] . \qquad (7)$$

From Theorem II it is evident that

$$s_{max} = \max_{X}\left[\frac{X_t\Gamma X}{X_tGX}\right] , \quad \text{and} \quad s_{min} = \min_{X}\left[\frac{X_t\Gamma X}{X_tGX}\right] .$$

Substitution of these quantities in (7), and subsequent multiplication by $g_{jj} \pm 2g_{jk} + g_{kk} \geq 0$, establishes the theorem.

Theorem IV: If a resistor having conductance $g_{jk} > 0$ joins the j^{th} and k^{th} nodes of S_i, but no inductor joins these nodes, then the conductance g_{jk} satisfies the inequality

$$g_{jj} + g_{kk} \geq 2\,\frac{|s_{max}| + |s_{min}|}{|s_{max}| - |s_{min}|}\, g_{jk} . \qquad (8)$$

Proof: The left-hand inequality in (5), taken with the negative option, reads

$$|s_{max}|(g_{jj} - 2g_{jk} + g_{kk}) \geq \gamma_{jj} - 2\gamma_{jk} + \gamma_{kk}. \qquad (9)$$

The right-hand inequality in (5), taken with the positive option, reads

$$\gamma_{jj} + 2\gamma_{jk} + \gamma_{kk} \geq |s_{min}|(g_{jj} + 2g_{jk} + g_{kk}). \qquad (10)$$

The absence of an inductor between the j^{th} and k^{th} nodes implies that $\gamma_{jk} = 0$. Under this condition, the left-hand member of (10) and the right-hand member of (9) are identical. Substitution of (10) into (9) and a slight rearrangement yields the desired result.

Theorem V: If S_i contains no parallel combination of a resistor and an inductor, then its natural frequencies suffer the separation

$$\left|\frac{s_{max}}{s_{min}}\right| \geq \frac{i + 1}{i - 1} \qquad (11)$$

Proof: Let S_i be as described in the theorem. In addition, assume that the largest conductance of S_i is connected between nodes α and β, and let this conductance be designated $g_{\alpha\beta} = g_{max}$.

By hypothesis, no inductor is in parallel with $g_{\alpha\beta}$. Thus the inequality of Theorem IV applies. This inequality reads

$$g_{\alpha\alpha} + g_{\beta\beta} \geq 2 \frac{|s_{max}| - |s_{min}|}{|s_{max}| + |s_{min}|} g_{max} . \qquad (12)$$

Now no more than i distinct conductances can emanate from the α^{th} node of S_i, and none of these is larger than g_{max}. Hence

$$i\, g_{max} \geq g_{\alpha\alpha} . \qquad (13)$$

A similar inequality applies to $g_{\beta\beta}$

$$i\, g_{max} \geq g_{\beta\beta} . \qquad (14)$$

Use of (13) and (14) in (12) yields

$$2i\, g_{max} \geq 2 \frac{|s_{max}| - |s_{min}|}{|s_{max}| + |s_{min}|} g_{max}. \qquad (15)$$

Cancellation of g_{max} from (15) and solution for the ratio $|s_{max}|/|s_{min}|$ yields the desired result, namely

$$\left|\frac{s_{max}}{s_{min}}\right| \geq \frac{i + 1}{i - 1}$$

Q.E.D.

From this point on, the development parallels that of Reference (2). Specifically by reasoning as in the proofs of Theorems 5, 6 and 7 of (2), one can deduce the following theorems.

Theorem VI: If S_i is good, then it must contain a parallel resistor-inductor combination.

Theorem VII: If S_i contains a parallel resistor-inductor pair, then the structure that results when the pair is short circuited is a Type I structure of complexity $i - 1$.

Theorem VIII: Let S_i contain a parallel resistor-inductor pair, and let S_{i-1} denote the Type I structure that results when the pair is short circuited. Then the natural frequencies of S_i and S_{i-1} satisfy the relationship

$$\left|\frac{s_{max}}{s_{min}}\right|_{S_i} \geq \left|\frac{s_{max}}{s_{min}}\right|_{S_{i-1}}$$

Next, by reasoning as in the proof of Theorem 9 of (2), one establishes

Theorem IX: If S_i is good, then S_i must be shunt-reducible (and thus loopwise reducible).

Finally, by arguing as in the proof of Theorem 10 of (2), one concludes that

Theorem X: S_i is good if it is loopwise reducible.

Thus the good Type I structures are those which are loopwise reducible; the bad Type I structures are those which are not.

TYPE II STRUCTURES

By carrying out an essentially dual development, one can arrive at the separation rule for Type II structures. The rule is as follows:

Theorem XI: A Type II structure is good if and only if it is series reducible (i.e., cut setwise reducible).

The only difficulty in establishing Theorem XI is that the nodal admittance equations do not have a counterpart in the case of nonplanar networks. This difficulty can be overcome, however, by using the following theorems in place of Theorems III, IV and V.

Theorem XII: Let j and k denote any two loops of a Type II structure. Then the resistances and the inductances contained in the two loops are such that

$$|s_{max}|(\ell_{jj} \pm 2\ell_{jk} + \ell_{kk}) \geq r_{jj} \pm 2r_{jk} + r_{kk} \geq |s_{min}|(\ell_{jj} \pm 2\ell_{jk} + \ell_{kk}) \qquad (16)$$

where

ℓ_{jj} = the total inductance in the j^{th} loop,

ℓ_{jk} = the inductance common to loops j and k,

r_{jj} = the total resistance in the j^{th} loop,

r_{jk} = the resistance common to loops j and k.

Proof: Identical to that of Theorem III, except that loop impedance matrices are used in place of the nodal admittance matrices, the loop currents i_j and i_k being chosen so as to traverse loops j and k.

Theorem XIII: If the intersection of loops j and k of a Type II structure contains inductance but no resistance, then the following relationship obtains:

$$\ell_{jj} + \ell_{kk} \geq 2 \frac{|s_{max}| - |s_{min}|}{|s_{max}| + |s_{min}|} \ell_{jk} .$$

Proof: Analogous to that of Theorem IV.

Theorem XIV: If a Type II structure contains no cut sets of the type shown in Fig. 7 of (2), then

$$\left| \frac{s_{max}}{s_{min}} \right| \geq \frac{i + 1}{i - 1}$$

where i denotes the complexity of the structure.

Proof: Every inductor of a Type II structure participates in a cut set, the remaining branches of which are resistors (for otherwise the inductor would be part of an inductor loop). If no cut sets of the type shown in Fig. 7 of (2) are present, then all such "mixed" cut sets must contain three or more branches. Thus if ℓ_{max} denotes the largest inductor in the structure, there must exist two distinct loops which intersect only in ℓ_{max}. Application of Theorem XIII to this pair of loops, and subsequent reasoning as in the proof of Theorem V, leads to the desired result.

The remainder of the derivation of Theorem XI parallels that of Theorems IX and X exactly.

DISCUSSION

A review of the results shows that the natural frequencies of all bad Type I and Type II structures suffer the constraint

$$\left| \frac{s_{max}}{s_{min}} \right| \geq \frac{n + 1}{n - 1} \tag{17}$$

where n denotes the structure complexity. This situation contrasts with that for the complementary tree structures, wherein the natural frequencies of bad structures satisfy

$$\left| \frac{s_{max}}{s_{min}} \right| \geq 4. \tag{18}$$

This fact naturally raises the question: Are the natural frequencies of bad Type I and Type II structures less severely constrained than those of bad complementary tree structures? Unfortunately we cannot answer this question at the present time. A computer-aided study of the natural frequencies of bad structures currently is in progress, however (4). Thus an answer to this and related questions should be available in the near future.

In our previous paper, we not only derived rules for distinguishing the good complementary tree structures from the bad, but we also made two conjectures regarding analogous rules for arbitrary two element kind structures. The results of the present paper confirm these conjectures for the case of Type I and Type II structures. Thus the conjectures remain unresolved only for two element kind structures which contain both one element kind loops and one element kind cut sets. It is hoped that this case will be settled in the near future.

FOOTNOTES

[*] This work was supported in part by the Joint Services Electronics Program under Contract DA 36-039-AMC-03200(E).

[1] Lee, H. B., "On the Differing Abilities of RL Structures to Realize Natural Frequencies", Proceedings of the Second Annual Allerton Conference on Circuit and System Theory; 1964.

[2] Lee, H. B., "On the Differing Abilities of RL Structures to Realize Natural Frequencies", IEEE Transactions on Circuit Theory; September, 1965.

[3] Hildebrand, F. B., Methods of Applied Mathematics, Prentice Hall, Inc., Englewood Cliffs, N.J., 1952; Chap. I, problems 76, 77 and 78.

[4] Murphy, P. J., S.M. Thesis, Massachusetts Institute of Technology.

OUTER AND INNER BOUNDS ON DRIVING-POINT IMPEDANCE FUNCTIONS OF LLFP NETWORKS*

T. S. Huang
Department of Electrical Engineering and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

W. H. Ku
Applied Research Laboratory
Sylvania Electronic Systems
Waltham, Massachusetts

ABSTRACT

I. Introduction

Recently, Huang and Lee[1] established certain outer bounds in the Z(impedance)-plane within which the impedance $Z(j\omega)$ of an R, ±L, ±C network must lie. In particular, they proved that the driving point impedance Z(s) satisfies the inequality

$$\left|Z(j\omega) - \frac{1}{2}(R_o+R_s)\right| \leq \frac{1}{2}(R_o-R_s) \quad , \quad \text{for all real } \omega \; ; \qquad (1)$$

where

$$R_o = Z(s)\Big|_{\text{All reactances open-circuited}} \; ,$$

$$R_s = Z(s)\Big|_{\text{All reactances short-circuited}} \; .$$

---

*The work of the first-named author was supported in part through the Joint Services Electronics Program (Contract DA36-039-AMC-03200(E)) by the U.S. Army Research Office, Durham; and in part by National Science Foundation (Grant GP-2495), the National Institutes of Health (Grant NH-04737-05), and the National Aeronautics and Space Administration (Grant NsG-496).

The purpose of the present paper is two-fold. First, we point out that inequality (1) applies to all LLFP (linear, lumped, finite, and passive) networks, not necessarily bilateral, if we interpret R_o and R_s appropriately. A proof of this generalization is given, which makes use of matrix algebra and is completely different from and more direct than the original proof in Ref. 1, which employed energy argument. Secondly, some inner bounds are established for driving-point impedance functions of networks containing positive resistances and two reactances. Inner bounds are also presented for certain special types of "R, n-reactance" networks, which indicates that a similar relation may exist for general "R, n-reactance" networks.

II. Outer Bounds for LLFP Driving-Point Impedance Functions

A generalization of inequality (1) is stated as:

Theorem 1. Let Z(s) be the driving-point impedance function of a network N containing R, ±L, ±C, and gyrators. Then

$$\left|Z(j\omega) - \tfrac{1}{2}(R_o+R_s)\right| \leq \tfrac{1}{2}(R_o-R_s) \quad ; \quad \text{for all real } \omega \qquad (2)$$

where

$$R_o = Z(s)\Big|_{\substack{\text{All reactances and gyrators with real}\\ \text{constants open-circuited}}}$$

and

$$R_s = Z(s)\Big|_{\substack{\text{All reactances and gyrators with real}\\ \text{constants short-circuited}}}$$

Theorem 1 is proved by first extracting all the reactances and gyrators from the network N and expressing the driving-point impedance (or, equivalently, the input scattering coefficient) in terms of the resistive n-port parameters and the reactive terminations. Then the inequality given by (2) is derived by invoking some well-known lemmas in matrix algebra.

III. Inner Bounds for "R, 2-Reactance" Driving-Point Impedance Functions

The main result is stated as:

Theorem 2. Let Z(s) be the driving-point impedance of a network containing positive resistances and two reactances which may be either positive or negative. Then

$$\left|Z(j\omega) - \frac{1}{2}(R_{os}+R_{so})\right| \geq \left|\frac{1}{2}(R_{os}-R_{so})\right| \quad ; \tag{3}$$

for all real ω

where

$$R_{os} = Z(s)\Bigg|_{\substack{\text{Reactance No. 1 open-circuited and}\\ \text{reactance No. 2 short-circuited}}}$$

and

$$R_{so} = Z(s)\Bigg|_{\substack{\text{Reactance No. 1 short-circuited and}\\ \text{reactance No. 2 open-circuited.}}}$$

IV. Inner Bounds "R, n-Reactance" Driving-Point Impedance Functions

Inner bounds were obtained for certain special types of "R, n-Reactance" networks.

Theorem 3. Let Z(s) be the driving-point impedance of a network which is a series combination of n parallel RC and parallel RL subnetworks, and let the resistances be $R_1 > R_2 > \cdots > R_n \geq 0$. Then

$$\left| Z(j\omega) - \frac{1}{2}\left[R_1 + \sum_{k=2}^{n} R_k\right]\right| \geq \frac{1}{2}\left[R_1 - \sum_{k=2}^{n} R_k\right] \quad ; \qquad (4)$$

for all real ω

It is noted that inequality (4) gives no useful inner bounds, when

$$R_1 \leq \sum_{k=2}^{n} R_k \quad .$$

Notice also that for the case of n=2, inequality (4) reduces to a special case of inequality (3).

Theorems 2 and 3 indicate that there may exist an inequality similar to (3) for general "R, n-reactance" networks.

REFERENCE

1. T. S. Huang and H. B. Lee, "Bounds on Impedance Functions of R, ±L, ±C Networks," Journal of the Franklin Institute, February 1965.

NON-LINEAR CIRCUIT MODELING

JOSEPH R. EDENHOFER
Autonetics
Anaheim, California

ABSTRACT

In the past, verification of electronic design was accomplished by "bread-boarding" a circuit to determine if it performed properly; however, this type of verification is no longer sufficient. The demands of increased reliability require that a circuit be tested in a manner which insures that the affect of time and environment will not have a degrading effect on circuit behavior. In this respect, computers have become an invaluable aid for circuit analysis. Through a knowledge of the drift and environmental characteristics of the circuit components, it is possible to determine the circuit's behavior under all possible conditions before the circuit is constructed.

This type of analysis has been accomplished by using node voltage or loop current matrix equations to represent the circuit. In performing this analysis each node voltage or loop current is determined. For most circuits all of these variables are not desired and represent unnecessary computation. For this reason as well as the introduction of integrated circuits and the increased emphasis on standard circuits, a circuit modeling technique which carries through the minimum number of variables is desirable.

The primary objective of this paper is to demonstrate a circuit-modeling technique based on the concept of state variables of system theory coupled with the theory of the indefinite admittance matrix. This circuit model will completely characterize a non-linear circuit with the minimum number of output variables. Once the circuit models are derived, they will be used to determine the electrical response of a system of interconnected circuits regardless of their interconnection. Since the circuit model describes the circuit regardless of its termination, the model need only be derived once.

This paper is concerned with the two steps which must be taken to implement the non-linear circuit modeling concept. The first step is to derive the circuit model from a description of the circuit topology. The second step consists of finding the transient or steady-state response of interconnected circuits described by the circuit models and their interconnection.

AN APPLICATION OF THE 4-COLOR "THEOREM" TO NETWORK THEORY

LEE M. MAXWELL, PH.D. and RICHARD E. SAEKS, M.S.E.E.
Department of Electrical Engineering
Colorado State University
Fort Collins, Colorado

INTRODUCTION

Although various techniques have been developed for determining whether or not a given graph is planar [5,6], the application of ideas associated with the 4-color "theorem" (yet unproven in general) to this purpose has not been extensive. The 4-color "theorem" has been proven for graphs containing less than 38 vertices [4,8] and is usually accepted as valid for the general case. Results based on the 4-color "theorem" are at least valid for graphs of less than 38 vertices. In this paper, equivalent statements to the 4-color "theorem" are given, which characterize the seg and the circ* matrices of planar graphs and which yield a necessary condition for a graph to be planar. Also resulting from the discussion of the 4-color "theorem" is a technique of detecting non-planar graphs which is readily amenable to digital computer manipulation.

In the procedure of testing for non-planar graphs developed here, it is necessary to show the existance of two segs, the union of which covers the graph. An algebraic method of identifying these two segs is described and generalized to a method of enumerating all of the segs or unions of segs which contain only a specified set of elements of the graph and/or do not contain another specified set of elements.

THE 4-COLOR "THEOREM" AND AN APPLICATION

In the development, two new definitions and one not normally used in graph theory are employed.

Definition 1. The smallest number of colors required to color all the vertices of a graph such that no element is incident to two vertices of the same color is the chromatic index of the graph. [1]

Definition 2. Let the vertices of a graph be segregated into exactly k mutually exclusive and all inclusive vertex subsets. The sub-graph composed of the set of elements, each of which is incident to two different subsets, is a k-seg. If one or more of the vertex subsets is void, the k-seg is an improper k-seg. Remark: The 2-seg is the usual seg. [2]

Definition 3. A set of sub-graphs covers a graph if the union of this set of sub-graphs contains the graph.

---

* A circuit or an element disjoint union of circuits. [7]

While the 4-color "theorem" is usually stated in terms of region coloring, i.e., chromatic number [1], the following is an equivalent statement:

"Theorem" 1. Every planar graph has a chromatic index less than or equal to 4. [1]

There are three statements which are here proven to be equivalent to the 4-color "theorem". They are as follows:

S 1. Every planar graph is covered by some one of its 4-segs.
S 2. Every planar graph is covered by some two of its 2-segs.
S 3. Every non-separable planar graph is covered by some two of its circs.

Lemma 1. Let S and C be a seg and circ respectively of a planar non-separable graph G and let G' be the dual of G. The set of elements of G' corresponding to the elements of S form a circ in G', and the set of elements in G' corresponding to the elements of C form a seg in G'.

Proof: Let G be a planar non-separable graph and G' its dual. Every seg S in G is the symmetric difference (∇) of stars (7), i.e., $S = A_1 \nabla A_2 \ldots \nabla A_n$, where A_i is the star of the i^th vertex in G. The set of elements C_i in G' corresponding to the elements of A_i for a circuit. This is illustrated in Fig. 1.

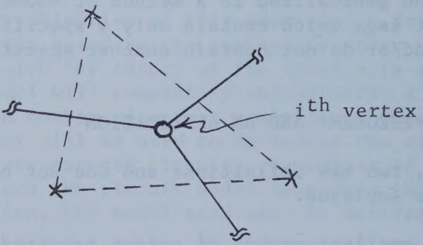


Fig. 1 Dual of a Star

Since there is a one-to-one correspondence between the elements of the planar graph and its dual, the dual (d) of the symmetric differences corresponds to the symmetric difference of the duals, i.e., $d(A_i \nabla A_j \ldots \nabla A_n)$ equals $C_i \nabla C_j \ldots \nabla C_n$, hence the set of elements in G' corresponding to the elements of S are symmetric difference of circuits, a circ, in G' [7]. Conversely a circ of G maps into a seg in G'.

"Theorem" 2. The four-color "theorem" and statements 1, 2, and 3 are equivalent.

Proof: (1) Four-color "theorem" if and only if S 1: The proof follows from the definition of a k seg and that a one-to-one correspondence can be made to exist between the k colors required to color the vertices of a planar graph, where k is less than or equal to 4, and the k segregations into which the vertices are put when forming the k-seg. If every element of the graph is incident to two vertices of different colors, then every element is incident to two vertices in different segregations of the k-seg, k less than or equal 4, and conversely.

(2) S 1 if and only if S 2. By S 2 some two 2-segs (segs), S_a and S_b, cover the graph. The vertex segregation for each of the segs are denoted a and a' and b and b' respectively. Label all the vertices of the graph by a or a' then by b or b'. Each vertex is then identified by either ab, a'b, ab', or a'b' and every element of the graph is incident to two vertices which are identified differently. Making the correspondence $c_1 = ab$, $c_2 = a'b$, $c_3 = ab'$, and $c_4 = a'b'$, S 1 follows. Conversely if S 1 is assumed, a like correspondence can be given and S 2 follows.

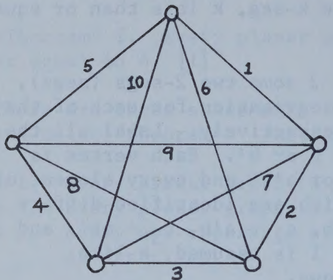
(3) S 2 if and only if S 3:

(a) If S 2 then S 3: let G be a non-separable planar graph. G', the dual of G [3], is planar and contains no 1-circuits with less than 2 elements. By S 2, G' is covered by some two of its 2-segs, say S_a and S_b. By lemma 1, C_a and C_b in G are circs. Since there is a one-to-one correspondence between the elements of the graph and its dual, C_a and C_b cover G.

(b) Every planar graph G is the union of its non-separable components. By an argument similar to 3(a) each of the non-separable components of G is covered by some two 2-segs. By permutting the vertex segregations of the non-separable components these 2-segs can be combined when necessary into a particular pair of 2-segs which completely cover G.

The applications of this theorem are twofold. First, the validity of the four-color "theorem" can be postulated thus permitting statements 2 and 3 to be used to characterize the seg and circ matrices of planar graphs. Secondly, an independent proof of any of the statements would suffice as a proof of the four-color "theorem".

As an example of the application of S 2 it is shown that the five vertex complete graph is non-planar. It is only necessary to write the seg matrix (matrix of all segs) of the graph and inspect the rows to determine whether or not two segs cover the graph. For graphs in general these techniques are amenable to digital computer manipulation. In the case of the five vertex complete graph (Fig. 2), no two segs cover the graph and it is thus non-planar. It may be noted that this test is sufficient for a graph to be non-planar but not necessary, as some non-planar graphs are 4-chromatic and therefore covered by two segs.



	1	2	3	4	5	6	7	8	9	10
1	1	1	0	0	0	0	1	0	1	0
2	1	0	0	0	1	1	0	0	0	1
3	0	0	0	1	1	0	0	1	1	0
4	0	0	1	1	0	0	1	0	0	1
5	0	1	1	0	0	1	0	1	0	0
6	1	0	1	0	0	1	1	1	1	0
7	0	1	0	1	0	1	1	1	0	1
8	0	0	1	0	1	0	1	1	1	1
9	1	0	0	1	0	1	0	1	1	1
10	0	1	0	0	1	1	1	0	1	1
11	1	1	1	1	0	0	0	0	1	1
12	1	1	0	1	1	0	1	1	0	0
13	0	1	1	1	1	1	0	0	1	0
14	1	1	1	0	1	0	0	1	0	1
15	1	0	1	1	1	1	1	0	0	0
16	0	0	0	0	0	0	0	0	0	0

Fig. 2 Five Vertex Complete Graph and Its Seg Matrix

COVERING OR PARTIAL COVERING WITH SEGS

The following is a general procedure for enumerating from the incidence information, the segs or union of segs which contain only a specified set of elements of a graph and/or do not contain another specified set of elements.

For an element X (x_1, x_2), which is incident to vertices x_1 and x_2, to be in a seg, it is necessary that x_1 and x_2 be in two different subsets. If the two different vertex subsets are denoted by 1 and 0, the requirement that the element X belong to the seg can be stated in terms of the Boolean equation,

$$S = x_1 x_2' + x_1' x_2. \qquad (1)$$

If it is desired to have elements $X(x_1, x_2)$, $Y(y_1, y_2)$, and $Z(z_1, z_2)$ in a seg, the requirement is logically stated as,

$$S = (x_1 x_2' + x_1' x_2)(y_1 y_2' + y_1' y_2)(z_1 z_2' + z_1' z_2). \qquad (2)$$

If it is required that some element not belong to the seg, the corresponding term is simply negated; i.e., $(x_1 x_2' + x_1' x_2)' = (x_1 x_2 + x_1' x_2')$. In general terms, the specification for a seg containing some elements and not containing others can be logically stated as,

$$S = \prod_n (n_1 n'_2 + n_1' n_2) \prod_m (m_1 m_2 + m_1' m_2'), \qquad (3)$$

where the n elements are specified to belong to the seg, and m elements are not to belong, and any others are "don't care" elements. By reducing Eq. (3) to the standard sum form and then taking into account the don't care terms, segs satisfying the specified criteria are identified. Each term in the standard sum form of Eq. (3) identifies the vertex partition of a seg which satisfied the specified criteria. Since a complete interchange of the classification of the vertex subsets does not alter the seg each term and its hinderance in the standard sum are equivalent and therefore can be combined as either of the terms.

Example 1: For the graph of Fig. 2 find all segs which contain elements 1 and 2 and do not contain elements 7 and 3.

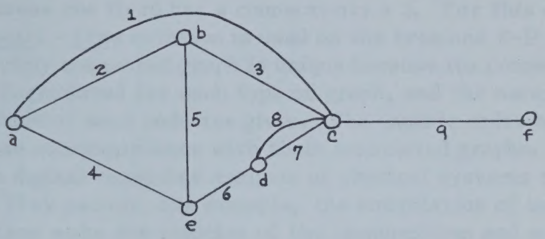


Fig. 2 Graph for Example 1

We write the logical statement:

$$\begin{aligned} S &= (ac' + a'c)(ab' + a'b)(bc + b'c)(cd + c'd') \\ &= (ab'c' + a'bc)(bcd + b'c'd') \\ &= (ab'c'd' + a'bcd). \end{aligned} \tag{4}$$

This can be reduced to only one term either ab'c'd' or a'bcd since one term is the hinderance of the other. By introducing the vertices not previously considered we obtain:

$$S = ab'c'd'ef + ab'c'd'e'f + abc'd'ef' + ab'c'd'e'f. \tag{5}$$

This result indicates that there are four segs satisfying the condition; i.e., the segs wherein vertex a is in one partition and vertices b, c, and d are in the other, whereas vertices e and f are in either.

If it is specified that the union of several segs will contain and/or not contain a set of elements the specification is formulated as:

$$S_u = \prod_n \Big[\sum_i^p (n_{1i}\, n'_{2i} + n'_{1i}\, n_{2i})\Big] \cdot \prod_m \Big[\prod_i^p (m_{1i}\, m_{2i} + m'_{1i}\, m_{2i}')\Big] \tag{6}$$

where p is the number of segs in the union. The problem of determining whether some union of two segs cover a graph can be formulated in the manner of Eq. (6). In searching for two segs which cover a graph use p = 2, and all of the elements contained in the graph are n elements. If, when Eq. (6) is expanded and found to be identically zero, it is concluded that there are no pair of segs which cover the graph.

While the logical algebra may become rather complex for a many vertex graph which is to be covered by several segs (v x s variables for a v vertex graph to be covered by s segs) computer programs are available for manipulating "many variable" logical functions. This then makes the technique a practical procedure for enumerating specified segs or determining whether a graph is non-planar from the incidence information.

References

1. C. Berge, The Theory of Graphs and Its Application, Wiley, New York, 1962.
2. M. B. Reed, "The Seg: A New Class of Subgraphs", IRE Transactions on Circuit Theory, CT 8, No. 1, (March 1961), pp. 17-22.
3. H. Whitney, "Non-Separable and Planar Graphs", Trans. Amer. Math. Soc., 34 (1932), pp. 339-362.
4. P. Franklin, "The Four Color Problem", American Journal of Mathematics, Vol. 44, (1922), pp. 225-236.
5. O. Wing, and G. J. Fisher, "An Algorithm for Testing Planar Graphs from the Incidence Matrix", 7th Midwest Symposium on Circuit Theory, University of Michigan, Ann Arbor, Mich.
6. P. M. Lin, "On Methods of Detecting Planar Graphs", Proc. 8th Midwest Symposium on Circuit Theory, Paper No. 1, Colorado State University, Fort Collins, Colorado, June 1965.
7. L. M. Maxwell and G. B. Reed, "Subgraph Identification - Segs, Circuits, and Paths", Proc. 8th Midwest Symposium on Circuit Theory, Paper No. 13, Colorado State University, Fort Collins, Colorado, June 1965.
8. P. Franklin, "The Four Color Problem", Scripta Math. Studies, 1961.

PLANE REPRESENTATIONS AND CODES FOR PLANAR GRAPHS[†]

LOUIS WEINBERG
Department of Electrical Engineering
City College of the City University of New York
New York, N. Y.

ABSTRACT

Unique plane representations of three basic types of planar graphs are derived, namely, the tree, the closed series-parallel (S-P) graph, and the triply connected graph. The first two types have connectivities of 1 and 2, respectively, whereas the third has a connectivity ≥ 3. For this derivation Polish or parenthesis - free notation is used on the tree and S-P graph, whereas a plane triply connected graph is unique because its connectivity ≥ 3. Simple codes are formulated for each type of graph, and the necessary and sufficient properties of each code are given. The canonic codes are unique, having a one-to-one correspondence with their associated graphs. The codes are very useful in digital-computer analysis of physical systems represented by linear graphs. They permit, for example, the compilation of unambiguous dictionaries and they make the solution of the isomorphism and automorphism group problems for linear graphs direct and simple. It is indicated that the general planar graph may be treated by modification or decomposition into graphs of the three basic types.

INTRODUCTION

Because of space restrictions we limit the discussion of applications and motivation for the codes; we also omit proofs of our theorems and assertions. We thus assume the reader is familiar with the necessity for using codes in the digital-computer analysis of systems represented by linear graphs. We further assume he is familiar with graph-theoretic concepts and definitions. We will give some definitions required for immediate understanding of the text; a suitable general reference is Ore's book[1]. Some other references[2,3] discuss isomorphism, 2-isomorphism, and automorphism groups of graphs, and give other useful definitions. The theory of triply connected (or 3-connected) graphs is presented in Tutte's paper[4]. An interesting application of triply connected graphs is given in Duijvestijn's thesis[5]. The references on series-parallel (S-P) and two-terminal S-P graphs are the Duffin paper[6], the Weinberg letter[7], and the Lawler and Salton paper[8]; the last-named paper also discusses Polish notation and the parenthesis-free codes in sufficient detail for our purposes.

We consider planar graphs, where by a planar graph is meant a graph that can be drawn in the plane without crossed branches. A planar graph that is so drawn is designated as a plane graph. A graph in which at most

---

[†] This research was jointly supported by the Radar Laboratory under Project Michigan (Department of the Army, Contract No. DA 28-043-AMC-00013(E)) and by the Institute of Science and Technology, The University of Michigan.

one branch joins each pair of nodes is designated as a simple graph; a graph with more than one branch incident to the same pair of nodes, that is, with parallel branches, is a multigraph.

We discuss the generation of codes for planar graphs, specifically, for three basic types, the tree, the S-P graph, and the triply connected graph, of connectivity 1, 2, and ≥ 3, respectively. Because of space considerations we do not consider the general planar graph. We merely mention here that a number of different methods may be used for generating a canonic code for such a graph, where we modify the graph or divide it into components in such a way that the algorithms for the three basic types may be used on the components.

First we discuss the generation of codes for triply connected graphs; we can refer to this code as the labyrinth-solving code or, more briefly, as the X-code, since we shall use X to represent the vector code. Then we consider trees for which Polish or parenthesis-free notation is used to generate a code; this is the Polish code or the parenthesis-free code or, more briefly, the T-code (T standing for tree). Polish notation in another form, where two different operators are defined, is used in codes for S-P graphs, which are discussed next; these are designated as L-codes (L standing for Lukasiewicz, who developed the notation), in order to distinguish them from the codes for trees. The S-P graphs we consider have connectivity 2, that is, every pair of branches in the S-P graph is part of a circuit.

As we shall see, the Polish notation permits the determination of a unique plane representation of a tree and of an S-P graph. This is the basic problem that must first be solved before a canonic code for a planar graph can be generated. After the unique representation is found, we can, if it is desirable, use the X-code on the tree and the S-P graph; for example, we may wish to compile a dictionary of trees, S-P graphs, and simple triply connected graphs using only one type of code. Alternatively, the Polish notation provides a canonic code for a tree and for an S-P graph.

The codes for the triply connected graph are new. Though the codes for the tree and the series-parallel graphs are known to switching theorists, the application of these codes to solve problems of interest, like drawing the series-parallel graph in a unique way so that a one-to-one correspondence exists between the canonic code and the graph, determining whether two series-parallel graphs are isomorphic or whether they are 2-isomorphic, determining whether two trees are isomorphic or whether one tree is isomorphic to a subtree of another tree, and finding the automorphism group of a tree or a series-parallel graph are original with the writer.

After explaining the codes we then briefly discuss how one or more dictionaries of planar graphs may be compiled. Such dictionaries have a variety of applications. They can be used, for example, to facilitate the search of patent files for previously used chemical formulas, where each formula is represented by a tree.

CODES FOR TRIPLY CONNECTED GRAPHS

In this section, unless the contrary is stated, the graphs considered are triply connected plane graphs.

We shall see that 4b codes are generated for a given graph where b is the number of branches. This set is necessary and sufficient because of a theorem[9] proved by the author elsewhere. This theorem states that the maximum order of the automorphism group of a planar triply connected graph is 4b. In fact, a more general condition may be stated, namely, that the order of the group, denoted by $|\Gamma|$, must satisfy

$$|\Gamma| \leq \min(2r\varphi_r, 2kp_k) \qquad (1)$$

where φ_r is the number of nodes of valence r, p_k is the number of meshes of length k, and the minimum is taken over all r and k, where r and k are positive integers taking on the values of the valences of the nodes of G and the lengths of the meshes of G, respectively.

Also, unlike planar graphs of lower connectivity, the triply connected graph has essentially a unique plane representation. By this is meant that the set of meshes of the graph is unique, where a mesh is a circuit enclosing a region that contains no branches of the graph; the mesh is given by its cyclically ordered set of branches, where the order has the same sense (clockwise or counterclockwise) for each mesh. For our purposes, it is convenient to consider an oriented plane so that a plane graph and its mirror-image graph are considered two different representations.

A code for G is generated by traversing an Euler path; thus our algorithm is basically a maze-solver. The procedure we use was formulated by Trémaux. To guarantee that each graph will yield an Euler circuit, we traverse each branch precisely once in each direction.

The rules are given below. We call a node old or new if it has or has not been reached previously, respectively. A branch being traversed in one direction is termed old if it has been previously traversed in the opposite direction, and new otherwise. In traversing a branch we go from an initial node to a terminal node. Choosing an initial branch and direction along this branch, we follow these rules:

1. When a new node is reached, on a branch, take (that is, leave the node on) the right-most branch of the set of other branches at the node.

2. When an old node is reached on a new branch, go back in the opposite direction on this same branch.

3. When an old node is reached on an old branch, leave the node on the right-most branch that has not previously been traversed in that direction (that is, with its initial node being the node just reached). Thus branches are traversed only once in each direction.

4. We record each traversed branch and the direction of traversal.

5. As we visit (or reach) a new node we label it with a number in natural order. Thus the initial node of the first branch is 1, its terminal

node is 2, and the next new node is labelled 3.

6. We form a code consisting of the numbers of the nodes visited in the order in which they are visited. The code is thus given by a vector of dimension $2b+1$.

After performing this process with each of the b branches as an initial branch and using both directions on each initial branch, we have a set of 2b vectors. Then repeating the above on the mirror-image graph yields an additional set of 2b vector codes. We then order the codes lexicographically (or, more briefly, lexically), that is, like the words in a dictionary; for these codes, whose components are the integers $1, 2, \ldots, n$, this is equivalent to arranging the codes as numbers ordered in increasing magnitude, where if $n \geq 10$, the code number is understood to be to the base $n+1$. The ordered codes are considered the rows of a code matrix M, whose dimension is thus $4b \times (2b+1)$. This code matrix is unique; in fact, there exists a one-to-one correspondence between a graph G and its code matrix M. It can therefore be shown that two graphs G_1 and G_2, whose respective code matrices are M_1 and M_2, are isomorphic if and only if $M_1 = M_2$.

The first row of M is defined as the canonic code, or where no confusion results, merely as the code, for graph G. There is a one-to-one correspondence between G and its canonic code. Thus we can show[2] that two graphs are isomorphic if and only if their canonic codes are equal.

If we are given a code for a graph G, we can by reversing the procedure for generating the code, draw a plane representation of G; this graph may have a different infinite face, which of course is immaterial, but it will have the same sense of the ordering of the branches around each face. Without drawing the graph, we can list the b branches of the undirected graph; from the code in Eq.(2) we find that the branches are given by (1, 2), (2, 3), (1, 3), (3, 4), (1, 4), (3, 5), and (2, 5), where the numbers of the two nodes of each branch have been written in increasing order. We can now also write the node-to-branch incidence matrix of the graph; in order to make the incidence matrix unique -- that is, to eliminate possible permutations of columns -- we make the columns correspond to a lexical ordering of the branches, where each branch is written with its smaller number first.

It is clear that a code has a number of distinguishing properties. We may, in fact, specify a set of necessary and sufficient conditions by which a vector may be checked to see whether it qualifies as a code. These are given for a simple graph by Theorem 1 below.

Theorem 1: Let G be a plane triply connected graph that is simple. A vector $X = (X_1, X_2, \ldots, X_p)$ is a code for G if and only if X satisfies the following conditions:

(1) The dimension of X is $2b+1$, where b is the number of branches in G; the numbers used in the code are $1, 2, \ldots, n$, where n is the number of nodes in G.

(2) If numbers occurring for the first time in the code are ordered in a sequence in accordance with their position in the code (reading from left to

right), then they must be in natural order 1, 2,...,n.

(3) There are b different ordered pairs of successive numbers X_iX_j in the code, where $X_i \neq X_j$, and corresponding to each of these there is a pair in reverse order X_jX_i.

(4) If the ordered pair of successive numbers X_iX_j occurs in the code and the number X_j has appeared previously but X_jX_i has not yet occurred, then the next number must be X_i; thus we obtain the triplet $X_iX_jX_i$. If X_jX_i has occurred previously or X_j has not appeared previously, the next number is $X_k \neq X_i$ so that we obtain $X_iX_jX_k$.

(5) For the first occurrence of the number X_j in the code, where $X_j = 2, 3, \ldots, n$, we note the immediately preceding number X_i. Now find X_jX_i in the code; then X_j cannot appear again in the code after this appearance of X_jX_i.

Though the above theorem is stated for a triply connected graph, it also holds for a biconnected plane graph. However, if the graph is not simple, some slight changes must be made which should be apparent to the reader. In condition (3), for example, the ordered pairs of successive numbers are no longer all different, and in condition (4) the triplet $X_iX_jX_i$ can occur even if X_j has not occurred previously or if X_jX_i has occurred previously. The changes for the 1-connected graph should also be clear; for example, in condition (4) the triplet $X_iX_jX_i$ will occur when X_j has unity valence and hence X_j cannot have occurred previously in the code.

There are, of course, other codes for plane triply connected graphs, some of which are much simpler to describe than our codes. However, they do not have the desirable properties possessed by our codes. For example, the code may be one of a very large number of codes (increasing greater than factorially with the number of nodes and meshes), each of which describes the graph; secondly it may be impossible to recognize whether the code corresponds to a graph without actually drawing the graph; thirdly and most important, a set of codes doesn't make evident the automorphism group of the graph, and a set of codes for each of two graphs doesn't automatically indicate whether the two graphs are isomorphic.

Such a simple code that may be mentioned is the mesh code devised by Duijvestijn[5]. The procedure for generating a code describes each mesh by the ordered node numbers on its contour. Given a code we can draw the corresponding graph by merely reversing the procedure.

We now consider an example to illustrate the procedure for generating codes. We generate a code for the graph shown in Fig. 1a, using branch a in the direction from left to right as the initial branch and direction. The cyclic path is shown in Fig. 1b for which the code is

$$X = 123134143525321 \tag{2}$$

In the same manner, using each branch in each of its directions as an initial branch and direction in this graph and its mirror-image graph, the reader can generate the complete set of 4b = 28 codes. He will obtain, for example, using branch e in a downward direction as the initial branch and direction, the code X(e) = 123132451525421. He will also find that some of these codes are equal. The significance of these equal codes is discussed in a companion paper[3].

CODES FOR TREES

The basic problem, as we indicated previously, is to find a representation in the plane that is a canonic one, that is, the plane graph that always has the same cyclic order of branches at each node. It is clear that there are many ways to draw a tree, the number of ways increasing exponentially with the number of branches. However, it is possible to obtain a canonic representation for a tree and thus a simple code can be generated, and the test for isomorphism of trees and the determination of the group of a tree become direct and simple. The method for doing this, which uses a parenthesis-free notation[†] developed by Lukasiewicz and is thus called Polish notation or parenthesis-free notation, is considered later. First we define some terms.

We use the terms center, bicenter, radius, distance, and diameter for a tree T, as defined by Ore. We also define any tree with a distinguished node as a rooted tree; the distinguished node is the root of the tree. It is often desirable to use the center C_o or one of the bicenters C_1 and C_2 as the root of a tree; as we shall see, this permits us to obtain a unique plane representation of a tree and a canonic code.

A procedure for finding the center or the central branch of a tree T is simple. We delete the nodes of unity valence (and of course their incident branches) to obtain a subtree T_1; such nodes are often called free nodes and the branches incident to them are called free branches. We repeat this process on T_1, removing the new free nodes to obtain T_2. The step is repeated until a single node or a single branch remains; the single node is the center C_o and will be obtained if the diameter δ is even. The single branch is the central branch whose nodes are the bicenters C_1 and C_2; it will be obtained when δ is odd.

A convenient representation of a rooted tree is given by drawing its nodes at consecutive levels; if the tree is not rooted, we select a root a_o and then use the following construction[1]. We place a_o and consider it the top of the tree at the zero level, designated by 0-level; this is the highest level. All branches incident to a_o are drawn to the nodes at a distance 1 from a_o; these nodes are lined up horizontally and are in the 1-level. From each of these nodes one draws the branches to the nodes at a distance 2 from a_o; these nodes are in the 2-level; and so on. Thus the numbering of the consecutive levels corresponds to the distance from a_o. From a node a_n at distance n from a_o there will be a single branch to a unique preceding node a_{n-1} of distance n - 1 and also a certain set of branches from a_n to a_{n+1} of distance n + 1. To none of these nodes a_n can there be any branches joining it to a node of the same or smaller distance except (a_{n-1}, a_n). If this branch is deleted, the tree is divided into two parts, one of which is a subtree containing the node a_n; this subtree can be considered a rooted subtree with root a_n.

[†] The writer is indebted to his former colleague Dr. R. M. Karp for introducing him to Polish notation.

The tree whose construction we have described is of course an undirected tree, but it is convenient for the purpose of determining the same unique weight for a node both when it is a part of the complete tree and when it is considered as a rooted subtree formed by deleting a branch as indicated above, to assume that the branches are directed away from the higher level. We then assign a weight equal to 1 - outward valence of the node. Thus for the rooted tree shown in Fig. 2a, where levels have been indicated and each node is given an identifying letter, the weights of the nodes are given in Fig. 2b.

Now we can use the Polish notation[8] for obtaining the code of any rooted tree. The codes correspond to the weights of the operators and variables used in the alphabet of the reference. The procedure is recursive, starting with the subtrees at the farthest distance from the root of the tree. The first number of the code is the weight of the root of the subtree, and the next numbers are the codes of the nodes of the subtree at a distance 1 from the root, ordered in correspondence with the nodes considered from left to right. Thus the code for subtree with root c (which, for brevity, we can call subtree c) is -31111 and that for subtree e is -111. We then go up to the next level. Subtree b has two branches at unit distance, one branch terminating in a node with code 1 and the other branch in node c with code -31111; thus the code for subtree b is given by -11-31111. Similarly, the code for the subtree rooted at d is -211-111. All the subtrees with their codes and the complete tree are shown in Fig. 3.

Now the codes of the subtrees can be used for redrawing the rooted tree so that there will be a one-to-one correspondence between a rooted tree and its code, which can thus be called the canonic code. We use lexical ordering where a negative number is considered to precede a positive one in the alphabet, and redraw the tree so that its code will be the earliest possible entry in a code dictionary. We do this in the same way as we determined the code, starting from the bottom level and working up; the nodes at the same level in a subtree are arranged from left to right in lexical order. Thus subtrees c and e remain the same and hence their codes do not change; however, subtree b has its branches interchanged, and subtree d is redrawn so that node e is on the left. The new codes for subtrees b and d are determined; the complete tree with root a must have nodes b and d interchanged; its new code can now be determined. The new subtrees and tree with lexical ordering and their codes are shown in Fig. 4. The canonic Polish code for the rooted tree is thus T = -1-2-11111-1-311111.

Any other rooted tree with the same canonic code will be isomorphic to this tree, where isomorphism of rooted trees requires in the one-to-one correspondence of nodes that the root of one tree correspond to the root of the other. Equality of codes is both necessary and sufficient for isomorphism.

If a tree is not rooted, the canonic code can still be obtained. If the tree has a center C_0, we use it as the root and determine the canonic code for this rooted tree. If the tree has a central branch, we use as the root whichever of the bicenters gives the lower lexical code; if the tree is symmetrical so that the rooted subtree C_1 is isomorphic to rooted subtree C_2, then the codes for C_1 as the root and for C_2 as the root will be the same and either may be used. The canonic code determined in this way is unique and is defined as the canonic code for the unrooted tree. The necessary and sufficient condition for isomorphism of two unrooted trees is that their canonic codes are equal.

In some problems it may be desirable to use the X-code for a tree. Since the Polish code gives a unique plane representation of the tree, we can apply the algorithm of the second section to the unique plane tree to obtain the 4b X-codes, and the complete discussion of the second section may now be repeated. For example, two rooted trees (or two unrooted trees) are isomorphic if and only if their code matrices are equal; and two trees are isomorphic if and only if their canonic X-codes are equal.

From the point of view of computing the canonic X-code, it should be pointed out that it can be determined without carrying out lexical ordering and even without determining all 4b codes. We choose the subtree, if it exists, that has the largest number of nodes of unit valence; we choose as the initial branch and direction the branches on the extreme left and the downward direction. Of course, it is also possible, if this is convenient, to choose a more easily recognized branch as the initial branch for generating a newly defined canonic X-code. For example, we can use the branch on the extreme left (not necessarily at the lowest level) of the tree as the initial branch and traverse it in the upward direction; this can be the canonic code by definition.

CODES FOR SERIES-PARALLEL GRAPHS

We now consider S-P graphs[6,7] of connectivity 2, which Duffin calls closed S-P graphs. We use the definition of a two-terminal graph and subgraph (that is, a doubly rooted graph, where a first root and a second root must be specified) of the Lawler and Salton reference[8], and assume the reader is acquainted with Polish notation, as explained in that reference. Thus he knows the necessary and sufficient conditions on a code, that is, he can determine whether it is well-formed, and he can determine the two-terminal subgraphs associated with an operator. Following their notation we use an alphabet consisting of the operators P_i and S_i and the variables $\{a, b, c, \ldots\}$. For any two-terminal S-P graph we will use the reduced Polish code as defined in the reference, that is, with operators of the maximum possible degree.

In order to obtain a unique plane representation of the graph, we adopt certain conventions. The first terminal (terminal 1, say, or terminal A) will be placed on the left and the second terminal (terminal 2 or terminal B) on the right. In writing the code for a graph containing two-terminal subgraphs between any pair of terminals, we consider the subgraphs in order from top to bottom; each of the two-terminal subgraphs is then considered from left to right. In this way the Polish code for any particular plane representation of a two-terminal S-P graph (where terminal 1 is on the left, terminal 2 is on the right, and the two-terminal subgraphs between terminal 1 and terminal 2 have a definite order from top to bottom) can be formulated. It is of course unique for this plane representation. However, to find the canonic code for the two-terminal S-P graph, it is clear that we must specify a canonic plane representation. In other words, we must draw the two-terminal subgraphs between terminal 1 and terminal 2, ordered from top to bottom, so that the lowest lexical code is obtained. This may be done by working directly with a plane representation of the two-terminal S-P graph or by working with a tree which is generated from the Polish code for the graph. The latter procedure is very efficient. Each rooted subtree and the tree itself have an operator as a root; its nodes at unit distance are its associated two-terminal branches or its associated operators. Since we can easily determine the two-terminal subgraphs associated with each operator, the corresponding rooted subtrees are made evident; the trivial subtrees, that is, those consisting of a single node, correspond to two-terminal subgraphs consisting of a single branch.

In the above we have spoken of the lowest lexical code; this is defined in terms of the weights of the operators and variables of the Polish code; for each character in the code its weight is substituted and we thus obtain the same type of code as was used previously for a tree. Codes of this type can therefore be ordered lexically.

What remains is to show how the canonic code is determined for a series-parallel graph without roots. As for the tree, where the rooted tree is utilized in defining the canonic code for an unrooted tree, so for an S-P graph, the canonic code is defined in terms of the codes of two-terminal S-P graphs formed from the given S-P graph. We merely consider all the two-terminal S-P graphs that can be formed from the graph, where only adjacent nodes can form a pair of terminals, that is, only those nodes that are the endpoints of a branch are a possible pair of terminals. We then find the lowest lexical code of all the possible two-terminal S-P graphs; this is defined as the canonic code for the graph without roots. Since we must consider a pair of nodes in each of two directions, it would seem that a maximum of 2b two-terminal graphs would have to be inspected. However, the number of graphs that must be inspected is very small. First, branches in parallel determine only one pair; secondly, as is shown in an example in the companion paper[3], considering the two-terminal graph in one direction automatically gives the lowest code for both directions. Most important, we recognize that the lowest code will correspond to a pair of terminals that has the largest number of subgraphs in parallel. Therefore, it is not necessary to consider a pair of nodes, either of which has a low valence. For the graph in the example of the reference[3] the nodes of branch g considered as terminals yield a graph with only two parallel subgraphs; so do the nodes of any other branch, either of whose nodes has valence 2; in fact, the maximum possible number of parallel subgraphs is equal to the lower of the two valences of the pair of nodes. By the above considerations we can often reduce the number of two-terminal graphs to be tested to the minimum possible, that is, one. This is illustrated in the detailed example of the companion paper[3]. It is also illustrated in that example how a unique rooted tree is made to correspond to a two-terminal S-P graph or to an S-P graph.

DICTIONARY OF GRAPHS

For many purposes it is desirable to have a dictionary of graphs of a certain type. This dictionary should be unambiguous and easy to consult; for example, there should be no synonyms, that is, each word must correspond to a different graph; there should be no homonyms, that is, no word is repeated in the dictionary. In other words, there should be a one-to-one correspondence between a code word and a graph. One use of a dictionary is in Duijvestijn's problem[5] of squaring the rectangle. Duijvestijn found it necessary to store not only the codes of triply connected graphs but also their identification numbers; since our codes completely identify the graphs, we need store merely the codes. Another use is as a catalogue of trees representing chemical formulas used in patents; this would expedite patent search.

We consider below each type of graph in turn.

(A) Triply Connected Graphs. A dictionary of all simple triply connected graphs represented by their canonic X-codes can be easily constructed and stored in a digital computer. Tutte has shown[4] how to generate every simple triply connected graph of b branches, or its dual, from a triply connected graph of b-1 branches. Thus the dictionary can be enlarged indefinitely. We can, for example, store the Duijvestijn mesh codes of the graphs in the dictionary with the maximum number of branches b_{max}; by Tutte's simple procedure[5] these may be used to generate the triply connected graphs with one more branch. If we also store the automorphism groups[2] of each of these graphs, it is possible to guarantee that no duplicates are generated; that is, only nonisomorphic graphs will be generated, and thus no unnecessary computation will be carried out on codes of duplicates.

We have mentioned that it is possible to test a given planar graph to determine whether it is triply connected. Another method is suggested by the dictionary. We merely replace each set of parallel branches by a simple branch and then generate the canonic code for the graph. We now look it up in the dictionary; it is triply connected if and only if it is present in the dictionary that goes at least up to codes of its dimension.

(B) Trees. We can form a dictionary of all rooted trees using the Polish code. Such a dictionary will automatacally answer such questions as the determination of the number of rooted trees for a specified number of nodes.

A dictionary of unrooted trees could be useful, for example, in determining the number of nonisomorphic tree classes in a complete graph on k nodes and the actual nonisomorphic trees; it could also be used to determine the actual nonisomorphic trees in any graph on k nodes.

(C) Series-Parallel Graphs. The graphs can be represented by their Polish weight codes and a dictionary can be made of all the rooted S-P graphs and all the unrooted S-P graphs.

We can, of course, consider the corresponding rooted tree for each S-P graph and form the dictionary of the X-codes of these trees. We can also use the X-codes directly on the canonic plane representation of the S-P graph.

(D) Triply Connected Graphs, Trees, and Series-Parallel Graphs. After the canonic plane representation of a tree and of an S-P graph has been found, we can generate X-codes for these plane graphs. Thus, using only this one type of code, we can compile a dictionary that contains all simple triply connected graphs, all trees, and all S-P graphs containing no more than a specified number of branches.

CONCLUSION

We have shown how to specify codes for planar graphs of three distinct types. The same methods can be used to characterize the general planar graph. These codes are useful in many computer applications of linear graphs. They permit, for example, a direct solution of the isomorphism and automorphism group problems. They also permit the compilation of dictionaries of graphs. Such a dictionary has a number of uses, one of which is in the problem of "squaring the rectangle"; another use is in a search of patent files. The codes are also useful for directed graphs[3].

REFERENCES

1. Ore, O., Theory of Graphs, Amer. Math. Soc. Colloquim Publication, vol. XXXVIII, Providence, Rhode Island, 1962.

2. Weinberg, L., "A Simple and Efficient Algorithm for Determining Isomorphism of Planar Triply Connected Graphs," submitted for publication.

3. Weinberg, L., "Automorphism Groups for Planar Graphs," presented at the third Allerton Conference on Circuit and System Theory, 1965; published in this issue.

4. Tutte, W.T., "A Theory of 3-Connected Graphs," Proc. Koninkl. Nederl. Akademie Van Wetenschappen-Amsterdam, Series A, vol. 64, no. 4, 1961.

5. Duijvestijn, A.J.W., "Electronic Computation Squared Rectangles," Philips Research Reports, vol. 17, December 1962, pp. 523-613 (Reprint R 466).

6. Duffin, R.J., "Topology of Series-Parallel Networks," J. Math. Anal. and Applic., vol. 10, 1965, pp. 303-318.

7. Weinberg, L., and P. Slepian, "Series-Parallel Networks," IRE Trans. PGCT, vol. CT-4, no. 3, Sept. 1957.

8. Lawler, E.L., and G.A. Salton, "The Use of Parenthesis-Free Notation for the Automatic Design of Switching Circuits," IRE Trans. PGEC, Sept. 1960.

9. Weinberg, L., "On the Maximum Order of the Automorphism Group of a Planar Triply Connected Graph," submitted for publication.

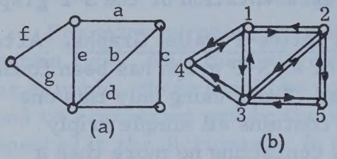

a
b
c
d
e
f
g
(a)
1
2
3
4
5
(b)

FIGURE 1

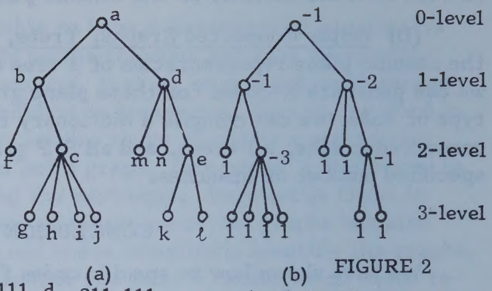

a
b
d
f
c
m
n
e
g
h
i
j
k
ℓ
(a)
-1
-1
-2
1
-3
1
1
-1
1
1
1
1
1
1
(b)
0-level
1-level
2-level
3-level

FIGURE 2

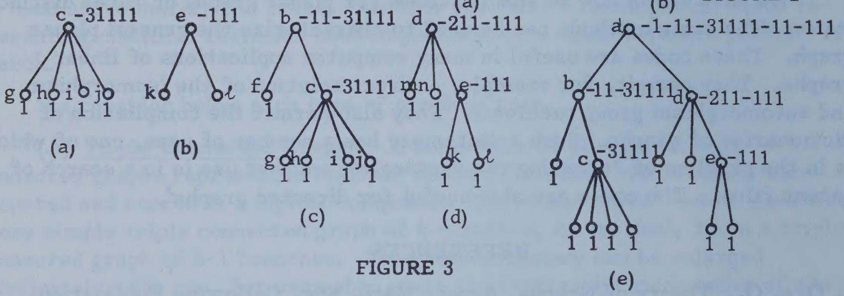

c -31111
g h i j
1 1 1 1
(a)
e -111
k ℓ
(b)
b -11-31111
f
1
c -31111
g h i j
1 1 1 1
(c)
d -211-111
m n
1 1
e -111
k ℓ
1 1
(d)
a -1-11-31111-11-111
b -11-31111
d -211-111
1
c -31111
1 1
e -111
1 1 1 1
1 1
(e)

FIGURE 3

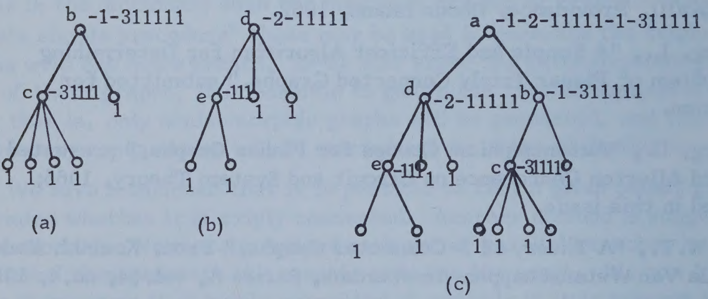

b -1-311111
c -31111
1
1 1 1 1
(a)
d -2-11111
e -111
1 1
1 1
(b)
a -1-2-11111-1-311111
d -2-11111
b -1-311111
e -111
1
1
c -31111
1
1
1
1 1 1 1
(c)

FIGURE 4

Fig. 1 Planar graph and a code-generating path that uses branch a from left to right as the initial branch and direction.

Fig. 2 A tree and weights assigned to each node

Fig. 3 The original tree and its subtrees with their codes.

Fig. 4 Subtrees and tree redrawn so that the codes have earliest lexical ordering.

AN ASYMPTOTE FOR LARGE SYSTEM MODELS

S. D. Bedrosian and J. Gulati
Moore School of Electrical Engineering Department of Physiology
University of Pennsylvania, Philadelphia, Pa. 19104

ABSTRACT

Linear graph models for very large systems have been approached via statistical techniques and by attempts to obtain formulas for asymptotes. The latter approach has produced such formulas for special categories of graphs. A brief review of some of this work by Otter, Renyi and Tutte is reviewed. Then the recently developed formula generating techniques are shown to lead to a general equation for an asymptote to the number of trees T for linear graphs with a large number of nodes and a minimum of constraints on configuration. This is illustrated and related to the previously published results.

INTRODUCTION

It has been noted [1] that the tree is probably the most important property of linear graph models for electrical network applications. The number and variety of problems utilizing linear graph models has been increasing. For some of these applications the exact number of trees is not required. In others an approximate number might be the crucial factor in determining whether or not the problem should be pursued further. Such considerations are important for complex problems in which N for the graph is very large. [2,3] Several attempts have been made for determining the asymptotic nature of certain other aspects of large graphs with limited success. For example, Renyi [4] demonstrated the existence of an asymptote for a class of graphs in which the nullity equals one. A similar formula has been obtained more recently by Tutte [5] which defines an asymptote for the number of rooted maps (with n edges) in planar graphs. Lupanov [2] gave a formulation for the number of non-isomorphic graphs having n edges.

DISCUSSION

A formula for finding the number of trees, T, in a complete graph of N nodes given by Cayley almost a century ago is

$$T = N^{N-2}.$$

For such a graph B = N(N-1)/2 branches. In recent years investigators [6,7,8] have found general formulas for computing T in other restricted classes of linear graphs. This sequence of events has culminated in a general technique [9] by which formulas can be generated for finding the number of trees in a graph provided its configuration is known. This is by no means a simple problem since there are a large variety of ways in which a given number of branches b can be removed from a given complete graph particularly for large N.

The objective here is to determine asymptotes for the number of trees in ordinary (non-oriented) graphs for $N \to \infty$. Such a result had been given by Otter [10]. His formulas were for nonhomeomorphic rooted trees and for rooted trees. The results given were for the special cases, degree of the nodes k = 4 related to the problem of isomers in chemistry, and $k = \infty$. The latter case yields

$$T_n \sim \frac{\beta^3 \alpha^{9/2}}{4 \sqrt{\pi}} \frac{\alpha^{-n}}{n^{5/2}} .$$

It is necessary that the degree of the nodes for the graph be known to evaluate the constants in this formula. Subsequently, it was shown by Harary [11] that Otter's results could be derived directly from Polya's theorem.

In the present approach the connected graphs are categorized in terms of the configuration of the subgraph or subgraphs of b branches deleted from the reference complete graph. A few trivial exceptions are made in order to retain the generality of the result obtained. The problem turns out to be quite straightforward when use is made of the formula generating technique mentioned earlier. This technique hinges on the definition of generic factors and recognition of the general form for expressions for the number of trees

$$T_{c-b} = \Pi \text{ (generic form)(Cayley's formula).}$$

Figure 1 gives a brief summary of generic factors for a number of special subgraphs given series designations. Figure 2 identifies four formula generating operations. [9,12] With these techniques it becomes possible to proliferate the number of such series. [13]. Figure 3 illustrates characteristic curves for graphs with a small value of N from which the designated configuration of branches has been deleted. Observe that the curves are already beginning to converge. The fact that the curves do not cross suggests the pattern of behavior for larger values of N. Utilizing judiciously selected factors and combinations of factors, numerical computations were made for large values of N in the range 100 to 1000. Regardless of the configurations used there is a distinct tendency for log T to converge.

It was found convenient to use Cayley's formulas as a reference. Using logarithms we find

$$\log T = (N-2) \log N.$$

Let

$$\log T = y \qquad \text{and} \qquad \log N = x.$$

Therefore,

$$y = (N-2)x$$

and for an asymptote, i.e., behavior of the graph as $N \to \infty$, we can assume $N - 2 \to N$. Thus for large N $y = Nx$. Taking the log of both sides again, we have

$$\log y = x + \log x.$$

It is easily shown that $x >> \log x$. Finally this converges to the straight line $\log y = x$. It turns out that a common asymptote exists for virtually all complete and incomplete graphs. It can be shown that the behavior persists even if branches are added rather than deleted from the reference complete graph. One of the special cases that is an exception to the rule is the simple polygon of N-nodes. In this case the number of trees is given by $T = N$. This yields, $\log y = \log x$. Note however that such a graph has a nullity of one. Renyi gave as the asymptote for graphs of N nodes and unit nullity

Number of graphs $C \sim \sqrt{\frac{\pi}{8}}\, N^{N-\frac{1}{2}}$

This result is found to have the same form as that shown above. One has but to take the log of both sides and as before note that the first term is very small compared to the second.

A general equation for an asymptote to the number of trees, T, of linear graphs for practically all configurations is given as

$$\log \log T = \log N + \log \log N.$$

REFERENCES

1. S. Seshu and M. B. Reed, *Linear Graphs and Electrical Networks*, Reading, Mass., Addison Wesley Publishing Co., Inc. 1961.
2. O. B. Lupanov, "Asymptotic Estimates of the Number of Graphs and Networks Having (n) Branches," Problems of Cybernetics, Vol. 4, pp. 1137-1155, Jan., 1963.
3. R. Kalaba, "Graph Theory and Automatic Control," Ch. 8 *Applied Combinatorial Mathematics*, E.F. Bechenbach, Ed., J. Wiley & Sons, N. Y., 1964.
4. A. Renyi, "On Connected Graphs, I.," Publ. Math. Inst. Hungarian Academy of Sciences, Vol. 5, pp. 385-388, 1960.
5. W. T. Tutte, " A Census of Planar Maps," Can. J. Math., Vol. XV, pp. 249-271, 1963.
6. L. Weinberg, "Number of Trees in a Graph," Proc. IRE, Vol. 46, No. 12, pp. 1954-1955, Dec., 1958.
7. P. V. O'Neil, "The Number of Trees in a Certain Network," presented Am. Math. Soc. Meeting, No. 604, Brooklyn, N.Y., Oct. 1963.
8. S. D. Bedrosian, "Formulas for the Number of Trees in a Network," IRE Trans. on Circuit Theory, Vol. CT-8, pp. 363-364, Sept., 1961.
9. S. D. Bedrosian, "Generating Formulas for the Number of Trees in a Graph," Jour. Frank. Inst., Vol. 277, No. 4, pp. 313-326, Apr.,1964.
10. R. Otter, "The Number of Trees," Ann. of Math., Vol. 49, pp. 583-599, 1948.
11. F. Harary, "Note on the Polya and Otter Formulas for Enumerating Trees," Michigan Math. J., Vol. 3, pp. 109-112, 1955-56.
12. Y. H. Ku and S. D. Bedrosian, "On Topological Approaches to Network Theory," Jour. Frank. Inst., Vol. 279, No. 1, pp. 11-21, Jan., 1965.
13. I. Taki, T. Kasai, S. Yoneda and H. Kusaka, "Mh-Series on the Total Number of Trees in Incomplete Graphs," Jour. Inst. Elect. Comm. Engr., Japan, Vol. 48, pp. 90-91, Jan. 1965.

Figure 1 STANDARDIZED SERIES FOR DELETING (OR ADDING) SETS OF BRANCHES FROM (TO) A COMPLETE GRAPH

Summary of standardized series to designate sets of branches in fixed configurations. In terms of generic factors [9] the p and r series [6] become

$$F_p = (N - 1)^{p-1}(N - 1 - p) \qquad F_r = (N - 2)^r$$

where

p relates to branches removed from a single specified node of G

r relates to branches removed from distinct pairs of nodes of G.

The corresponding formulas are,

$$T_{c-p} = F_p N^{N-2-p} \qquad T_{c-r} = F_r N^{N-2-r}.$$

The s and m series [8] are given in a systematic tabulation.

s	Generic Factors F_s
1	$(N-2)$
2	$(N-3)(N-1)$
3	$(N-2)(N^2 - 4N + 2)$
4	$(N^2 - 5N + 5)(N^2 - 3N + 1)$
5	$(N^2 - 4N + 3)(N^2 - 4N + 1)(N - 2)$
⋮	

m	Generic Factors F_m
3	$(N - 3)^2$
4	$(N - 2)^2(N - 4)$
5	$(N^2 - 5N + 5)^2$
6	$(N^2 - 4N + 3)^2(N - 4)$
7	$(N^3 - 7N^2 + 14N - 7)^2$
⋮	

The corresponding formulas become,

$$T_{c-s} = F_s N^{N-2-s} \qquad \text{for } N \geq s + 1$$

$$T_{c-m} = F_m N^{N-2-(m-1)} = F_m N^{N-1-m} \qquad \text{for } N \geq M \geq 3$$

The h series refers to a family of complete graphs with a hinged branch or branches incident at one node. The generic factor is

$$F_h = (N - N_c)^{N_c - 2}(N - 1)^h(N - N_c - h)$$

where

N_c = number of nodes in the complete graph being deleted

h = the number of ≥ 0 of hinged branches at the specified node.

The corresponding formula is

$$T_{c-h} = F_h N^{N-N_c-1-h}.$$

Figure 2

FORMULA GENERATING OPERATIONS

	SUBGRAPH(S) OF b BRANCHES	
	ADDED	DELETED
g	ADDITION $(+)$	DELETION $(-)$
$\bar{g}$	$*$ COMPLEMENTATION	$\&$ SUPPLEMENTARY COMPLEMENTATION

WHERE

$G = g \cup \bar{g}$ = COMPLETE GRAPH

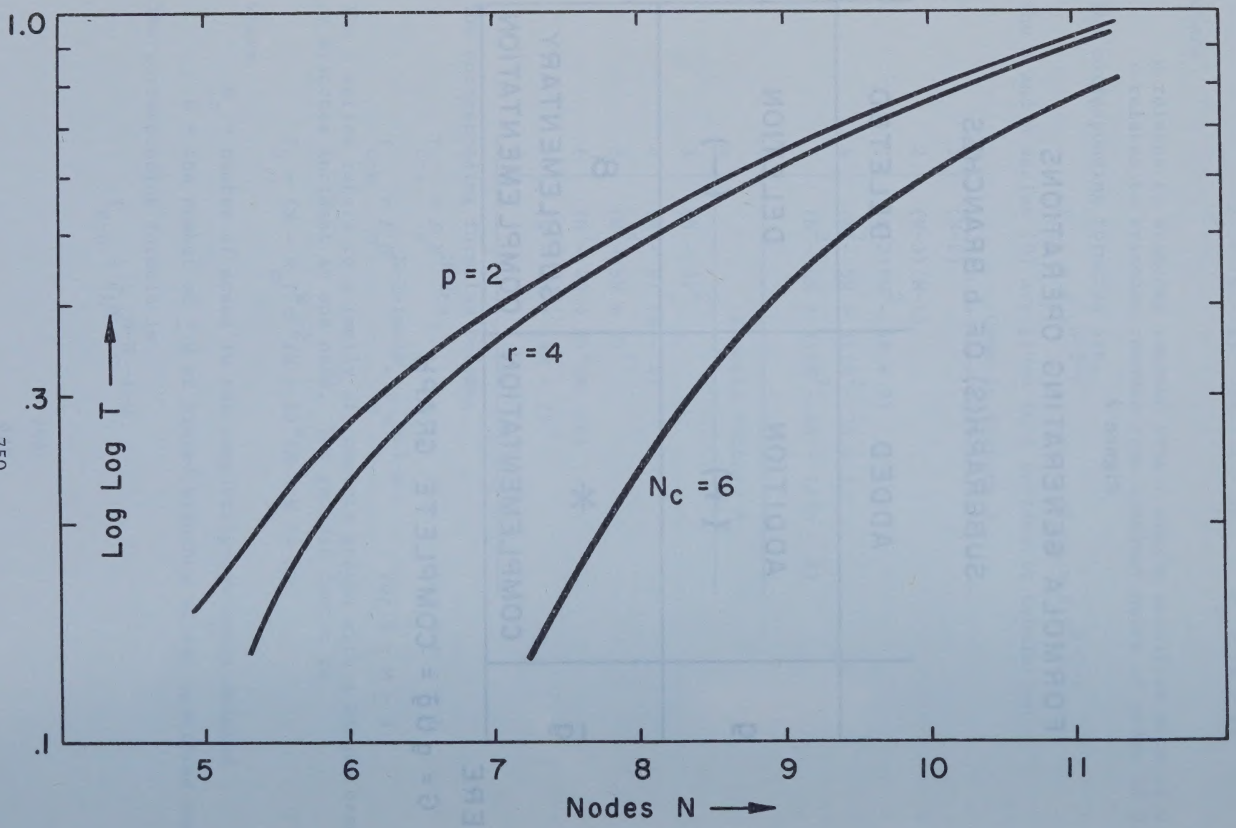


Figure 3 Curves of T for Incomplete Graphs vs Small Values of N.

ANALYSIS OF LOSSY COMMUNICATION NETS BY MODIFIED INCIDENCE MATRICES

TADAO MURATA
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois

ABSTRACT

In this paper two modified incidence matrices for more general nets are introduced; these reduce to the usual incidence matrix when all the branch flow efficiencies are unity. Vertex potentials for the generalized net are defined and its connection matrices are given. The use of the new incidence matrices has made it possible to obtain upper and lower bounds on the maximum flows between two vertices of a lossy communication net.

I. PRELIMINARIES

Recently Fujisawa [1] introduced the loss or efficiency factor on flow through the branch and considered the maximum flow problem of a lossy net using the maximality criteria in terms of vertex potentials. Mayeda and Van Valkenburg [2] derived many properties of lossy communication nets. Onaga [7] obtained further results. Also Jewell [8] considered a similar model. As yet no simply analytical method for lossy nets has been found. In this paper two new incidence matrices, which reduce to the usual incidence matrix when all the branch flow efficiencies are unity, are proposed for foundation of analysis of more general nets and a maximum flow problem in a lossy communication net is considered.

In many practical systems such as power distribution networks and other transmission systems, flow through the branch might be attenuated and could be amplified. It is, therefore, more practical and general to introduce the branch flow efficiency α such that the receiving flow of a branch $\underline{f}$ is equal to the sending flow of that branch $\overline{f}$ multiplied by α, i.e.

$$\underline{f} = \alpha \overline{f} \tag{1}$$

where α is a non-negative constant associated with each branch in a net. When $0 \leq \alpha < 1$, $\alpha = 1$ and $1 < \alpha$, flow is attenuated, conserved and amplified respectively as it passes through the branch. If the value of α is restricted in the range, $0 \leq \alpha < 1$, the value $\rho = 1 - \alpha$ is referred to as the loss factor and the net is said to be a lossy net, while the net with $\alpha = 1$ or $\rho = 0$ is said to be a lossless net. It is obvious that nets with the above flow property include the usual (lossless) net as a special case. In this sense, the nets considered here are more general. (From here on, the symbol N denotes the generalized oriented net(s) considered in this paper.)

II. MODIFIED INCIDENCE MATRICES AND CONSERVATION OF FLOW

The usual incidence matrix can not be applied to N because of the flow property. The modified incidence matrices are defined in the following definitions where α_j denotes the branch flow efficiency of branch j.

Definition 1: An alpha incidence matrix, $A_\alpha = [a_{ij}]$ of N with n vertices and m branches is an n by m matrix such that

$a_{ij} = 1$ if branch j has vertex i as an initial vertex

$a_{ij} = -\alpha_j$ if branch j has vertex i as a terminal vertex

and $a_{ij} = 0$ if branch j is not incident at vertex i.

Definition 2: An inverse alpha incidence matrix, $A_{1/\alpha} = [a'_{ij}]$ of N with n vertices and m branches is an n by m matrix such that

$a'_{ij} = 1/\alpha_j$ if branch j has vertex i as an initial vertex

$a'_{ij} = -1$ if branch j has vertex i as a terminal vertex

and $a'_{ij} = 0$ if branch j is not incident at vertex i.

Using A_α or $A_{1/\alpha}$, conservation of flow at each vertex of N which corresponds to Kirchhoff's current law in an electrical network can be expressed by means of matrices as follows:

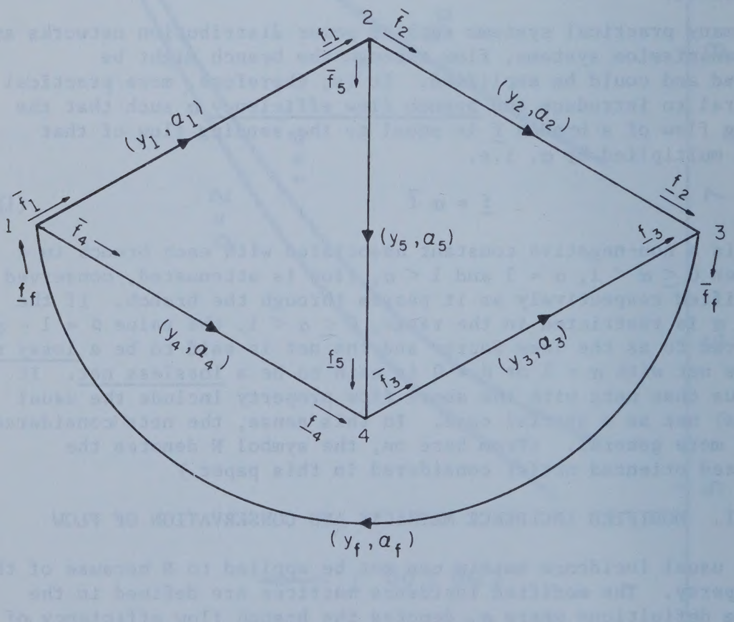
$$A_\alpha \overline{F} = 0 \tag{2}$$

or

$$A_{1/\alpha} \underline{F} = 0 \tag{3}$$

where $\overline{F}$ and $\underline{F}$ are column matrices of sending and receiving branch flows $\overline{f}$ and $\underline{f}$ respectively.

Example 1: Eqs. (2) and (3) corresponding to the net of Fig. 1 are given by Eqs. (2a) and (3a) respectively.



2
$\underline{f}_1$
$\overline{f}_2$
$\overline{f}_5$
(y_1, α_1)
(y_2, α_2)
1
$\overline{f}_1$
$\overline{f}_4$
$\underline{f}_2$
$\underline{f}_3$
3
(y_5, α_5)
$\underline{f}_f$
$\overline{f}_f$
(y_4, α_4)
$\underline{f}_5$
$\overline{f}_3$
(y_3, α_3)
$\underline{f}_4$
4
(y_f, α_f)

Fig. 1

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & -\alpha_f \\ -\alpha_1 & 1 & 0 & 0 & 1 & 0 \\ 0 & -\alpha_2 & -\alpha_3 & 0 & 0 & 1 \\ 0 & 0 & 1 & -\alpha_4 & -\alpha_5 & 0 \end{bmatrix} \begin{bmatrix} \overline{f}_1 \\ \overline{f}_2 \\ \overline{f}_3 \\ \overline{f}_4 \\ \overline{f}_5 \\ \overline{f}_f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{2a}$$

$$\begin{bmatrix} 1/\alpha_1 & 0 & 0 & 1/\alpha_4 & 0 & -1 \\ -1 & 1/\alpha_2 & 0 & 0 & 1/\alpha_5 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1/\alpha_f \\ 0 & 0 & 1/\alpha_3 & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \underline{f}_1 \\ \underline{f}_2 \\ \underline{f}_3 \\ \underline{f}_4 \\ \underline{f}_5 \\ \underline{f}_f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{3a}$$

III. VERTEX POTENTIALS (OR WEIGHTS)

One definition of a vertex potential for N is sufficient in order to obtain a set of upper and lower bounds on the maximum flows in N. However, the more sets of upper and lower bounds are computed independently, the better approximation for the maximum flows is obtained by choosing the least upper and greatest lower bounds from the sets of upper and lower bounds. Therefore, four different kinds of vertex potentials for N, which are independent of each other, are defined here and for each of them a set of upper and lower bounds on the maximum flows can be obtained through applying the procedures described later.

Let x_i be a vertex potential (or weight) of vertex i and let

$$x_i = \begin{cases} \overline{x}_i & \text{if vertex i is an initial vertex} \\ \underline{x}_i & \text{if vertex i is a terminal vertex.} \end{cases}$$

Let y_k denote a branch weight of branch k. y_k can be a branch (or edge) capacity C_k of branch k, branch flow efficiency α_k, the product of α_k and C_k or some other factor associated with branch k.

The first definition of a vertex potential for N is given by the relation between the sending branch flow $\overline{f}_k$ of branch k and the initial vertex potential $\overline{x}_i$ of branch k incident to vertex i, i.e.,

$$\overline{f}_k = y_k \overline{x}_i \ . \tag{4}$$

The second definition is given by the relation between $\overline{f}_k$ and the terminal vertex potential $\underline{x}_i$ of branch k incident to vertex i, i.e.,

$$\overline{f}_k = y_k \underline{x}_i \ . \tag{5}$$

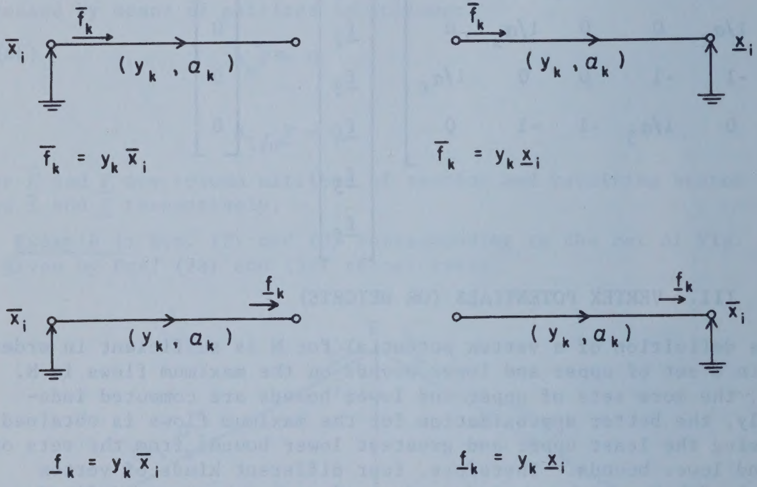
The third definition if given by the relation between the receiving branch flow $\underline{f}_k$ of branch k and $\overline{x}_i$, i.e.,

$$\underline{f}_k = y_k \overline{x}_i \ . \tag{6}$$

The last one is given by the relation between $\underline{f}_k$ and $\underline{x}_i$, i.e.,

$$\underline{f}_k = y_k \underline{x}_i \ . \tag{7}$$

Figs. 2-5 illustrates the above definitions of vertex potentials for N.



$\overline{f}_k$
$\overline{x}_i$
(y_k, α_k)
$\overline{f}_k = y_k \overline{x}_i$
$\overline{f}_k$
$\underline{x}_i$
(y_k, α_k)
$\overline{f}_k = y_k \underline{x}_i$
$\underline{f}_k$
$\overline{x}_i$
(y_k, α_k)
$\underline{f}_k = y_k \overline{x}_i$
$\underline{f}_k$
$\underline{x}_i$
(y_k, α_k)
$\underline{f}_k = y_k \underline{x}_i$

Figs. 2, 3, 4 and 5

Note that when $\alpha_k = 1$ for all k, the above four basic branch elements defined by Eqs. (4)-(7) reduce to the unistor [3], a branch-like element whose flow is proportional to one of its vertex potentials and independent of the other vertex potentials.

Eqs. (4)-(7) can be written in matrices as follows:

$$\overline{F} = Y(A^+)^T \overline{X} \tag{8}$$

$$\overline{F} = Y(A^-)^T \underline{X} \tag{9}$$

$$\underline{F} = Y(A^+)^T \overline{X} \tag{10}$$

and

$$\underline{F} = Y(A^-)^T \underline{X} \tag{11}$$

where Y is a diagonal matrix of branch weights y_k's, $\overline{X}$ and $\underline{X}$ are column matrices of vertex potentials $\overline{x}_i$ and $\underline{x}_i$ respectively. A^+ is the usual incidence matrix of a directed graph with all -1 entries replaced by zero and A^- is A with all +1 entries replaced by zero and all -1 entries replaced by +1.

IV. CONNECTION MATRICES OF N

The following matrix equations for N are obtained by multiplying Eqs. (8) and (9) by A_α and Eqs. (10) and (11) by $A_{1/\alpha}$ and by substituting Eq. (2) or (3) in the resulting equations.

$$A_\alpha Y(A^+)^T\overline{X} = 0 \tag{12}$$

$$A_\alpha Y(A^-)^T\underline{X} = 0 \tag{13}$$

$$A_{1/\alpha} Y(A^+)^T\overline{X} = 0 \tag{14}$$

and

$$A_{1/\alpha} Y(A^-)^T\underline{X} = 0 \ . \tag{15}$$

Observe the matrix products, $A_\alpha Y(A^+)^T$, $A_\alpha Y(A^-)^T$, $A_{1/\alpha} Y(A^+)^T$ and $A_{1/\alpha} Y(A^-)^T$ in Eqs. (12)-(15) which correspond to so-called connection matrices [4], [5]. Now we have the following four theorems for the connection matrices of N according to the four different vertex potentials defined in the previous section.

Theorem 1: Let a vertex potential for N be defined by Eq. (4). Then a connection matrix of N with n vertices, $A_\alpha Y(A^+)^T \equiv K^{(1)} = [k_{ij}^{(1)}]$ is an n by n matrix with rows and columns labeled in the same order by the n vertices and its entry is given by

$$k_{ij}^{(1)} = \begin{cases} \Sigma' y_k & \text{if } i = j \\ \Sigma''(-\alpha_k y_k) & \text{if } i \neq j \end{cases}$$

where the "summation" Σ' is over all the branches oriented from i, Σ" is over all the branches oriented from vertex j to vertex i and $k_{ij}^{(1)} = 0$ if there is no such branch.

Proof: Since Y is diagonal, $A_\alpha Y$ differs from A_α only in that the k^{th} column is multiplied by y_k. Observe that there are two nonzero entries in the j^{th} column of A_α. One is a 1 in the the i_1^{th} row where branch j is oriented from vertex i_1 and the other is a $-\alpha_j$ in the i_2^{th} row where branch j is oriented to vertex i_2. Now there is a nonzero entry y_k in the i_1^{th} row of $A_\alpha Y$ with a "1" in the corresponding position of the i_1^{th} column of $(A^+)^T$ for each branch k which is oriented from vertex i_1. Therefore the $i_1 i_1$-entry or the diagonal entry of $A_\alpha Y(A^+)^T$ is $k_{ii}^{(1)} = \Sigma' y_k$. There is also a nonzero entry $-\alpha_k y_k$ in the i_2^{th} row of $A_\alpha Y$ with a "1" in the corresponding position of the j^{th} column of $(A^+)^T$ for each branch k which is oriented from vertex j to vertex i_2. Hence the non-diagonal entry of $A_\alpha Y(A^+)^T$ is given by $k_{ij}^{(1)} = \Sigma''(-\alpha_k y_k)$ if there exists such a branch. Q.E.D.

Corresponding to the second, third and fourth definitions of a vertex potential, connection matrices for N are stated in the following theorems whose proofs are very much similar to one for Theorem 1. It is assumed in Theorems 2-4 that rows and columns are labeled in the same order by the n vertices as in Theorem 1.

Theorem 2: Let a vertex potential for N be defined by Eq. (5). Then a connection vertex of N with n vertices, $A_\alpha Y(A^-)^T \equiv K^{(2)} = [k_{ij}^{(2)}]$ is an n by n matrix and its entry is given by

$$k_{ij}^{(2)} = \begin{cases} \Sigma'''(-\alpha_k y_k) & \text{if } i = j \\ \Sigma''''y_k & \text{if } i \neq j \end{cases}$$

where the "summation" Σ''' is over all the branches oriented to vertex i, Σ'''' is over all the branches oriented from vertex i to vertex j and $k_{ij}^{(2)} = 0$ if there is no such branch.

Theorem 3: Let a vertex potential for N be defined by Eq. (6). Then a connection matrix of N with n vertices, $A_{1/\alpha}Y(A^+)^T \equiv K^{(3)} - [k_{ij}^{(3)}]$ is an n by n matrix and its entry is given by

$$k_{ij}^{(3)} = \begin{cases} \Sigma'(y_k/\alpha_k) & \text{if } i = j \\ \Sigma''(-y_k) & \text{if } i \neq j \end{cases}$$

where the "summations" Σ' and Σ'' are the same as in Theorem 1.

Theorem 4: Let a vertex potential for N be defined by Eq. (7). Then a connection matrix of N with n vertices, $A_{1/\alpha}Y(A^-)^T \equiv K^{(4)} = [k_{ij}^{(4)}]$ is an n by n matrix and its entry is given by

$$k_{ij}^{(4)} = \begin{cases} \Sigma'''(-y_k) & \text{if } i = j \\ \Sigma''''(y_k/\alpha_k) & \text{if } i \neq j \end{cases}$$

where the "summations" Σ''' and Σ'''' are the same as in Theorem 2.

Example 2: Although direct calculations of the matrix products, $A_\alpha Y(A^+)^T$ etc. give the connection matrices, they can be obtained by inspection from a graph of N through applying Theorems 1-4 respectively. For example $K^{(1)}$, $K^{(2)}$, $K^{(3)}$ and $K^{(4)}$ of the net in Fig. 1 are written by inspection as shown below.

$$K^{(1)} = \begin{bmatrix} y_1+y_4 & 0 & -\alpha_f y_f & 0 \\ -\alpha_1 y_1 & y_2+y_5 & 0 & 0 \\ 0 & -\alpha_2 y_2 & y_f & -\alpha_3 y_3 \\ -\alpha_4 y_4 & -\alpha_5 y_5 & 0 & y_3 \end{bmatrix}$$

$$K^{(2)} = \begin{bmatrix} -\alpha_f y_f & y_1 & 0 & y_4 \\ 0 & -\alpha_1 y_1 & y_2 & y_5 \\ y_f & 0 & -\alpha_2 y_2 - \alpha_3 y_3 & 0 \\ 0 & 0 & y_3 & -\alpha_4 y_4 - \alpha_5 y_5 \end{bmatrix}$$

$$K^{(3)} = \begin{bmatrix} y_1/\alpha_1 + y_4/\alpha_4 & 0 & -y_f & 0 \\ -y_1 & y_2/\alpha_2 + y_5/\alpha_5 & 0 & 0 \\ 0 & -y_2 & y_f/\alpha_f & -y_3 \\ -y_4 & -y_5 & 0 & y_3/\alpha_3 \end{bmatrix}$$

$$K^{(4)} = \begin{bmatrix} -y_f & y_1/\alpha_1 & 0 & y_4/\alpha_4 \\ 0 & -y_1 & y_2/\alpha_2 & y_5/\alpha_5 \\ y_f/\alpha_f & 0 & -y_2 - y_3 & 0 \\ 0 & 0 & y_3/\alpha_3 & -y_4 - y_5 \end{bmatrix} .$$

V. BOUNDS ON MAXIMUM FLOWS THROUGH LOSSY NETS

From here on, only the first definition of a vertex potential, Eq. (4) or Eq. (12) i.e.,

$$K^{(1)}\overline{X} = 0 \tag{12a}$$

is considered since slight modifications give parallel theorems for each of the other definitions of a vertex potential. Let branch weight y_k be branch (or edge) capacity C_k of branch k although y_k can take any other factor (e.g., α_k or $\alpha_k C_k$) associated with branch k. It is assumed that a lossy communication net to be analyzed has been modified in the following manner without affecting the maximum flows.

1) Eliminate all branches directed to the source and all the branches directed from the sink.
2) Add a branch directed from the sink to the source having the unknown branch efficiency α_f and normalized (unit) branch weight $y_f = 1$. This branch is called the feedback branch.

In order for the homogeneous equation (12a) to have non-trivial solutions for $\overline{X}$, the determinant of $K^{(1)}$ must equal zero. From this condition for non-trivial solutions (which gurantees existence of flow in a communication net) the unknown branch flow efficiency α_f will be uniquely determined. Observe that there is only one path from the sink to the source in the modified graph that is through the feedback branch. Therefore the meaning of the efficiency α_f is clear, i.e., α_f is the ratio of the total outgoing flow at the source to the total incoming flow at the sink ($\underline{f}_f = \alpha_f \bar{f}_f$) or $1/\alpha_f$ is an *equivalent efficiency* from the source to the sink in a lossy net under the condition of Eq. (4).

With the above value of α_f, Eq. (12a) can be solved for vertex potentials $x_1, x_2, \ldots, x_n$ within an arbitrary multiplicative constant by various methods, e.g. through an equivalent Coates' flow graph [5] of a lossy communication net which can be easily drawn using Theorem 1. Then an upper and a lower bounds on the maximum flows from source p to sink q in a lossy communication net can be derived as in the following theorems which are generalizations of the lossless case [6]. In the theorems to follow, $\bar{t}_{pq}$ denotes the *source terminal capacity* which is the total outgoing flow at source p to receive the maximum flow at sink q and $\underline{t}_{pq}$ the *sink terminal capacity* which is the maximum flow that can be received at sink q when the maximum flow is sent from p to q [2].

Theorem 5: Let x_{max} and x_{min} be the maximum and minimum vertex potentials in the set $\{\{x_1, x_2, \ldots, x_n\} - x_q\}$ respectively, where x_q is the vertex potential at sink q. Then

$$x_q/x_{max} \leq \underline{t}_{pq} \leq x_q/x_{min} \tag{16}$$

and

$$\alpha_f x_q/x_{max} \leq \bar{t}_{pq} \leq \alpha_f x_q/x_{min} \; . \tag{17}$$

Proof: Consider the first part of Eq. (16), $x_q/x_{max} \leq \underline{t}_{pq}$. When all vertex potentials are normalized by x_{max}, we have $\bar{x}_i \leq 1$ for all $i \neq q$. Then the definition of a vertex potential, $\bar{f}_i = C_i \bar{x}_i$ gives $\bar{f}_i \leq C_i$ for all $i \neq q$. It is easily seen also from the definition of a vertex potential that the outgoing flow at sink q is given by $\bar{f}_f = C_f x_q/x_{max} = x_q/x_{max}$ (recall that we put $C_f = 1$) under the normalization. Since there is only one path from sink q, the total incoming flow at q is the same as $\bar{f}_f = x_q/x_{max}$. In order to show that $\bar{f}_f \leq \underline{t}_{pq}$, suppose that $\bar{f}_f > \underline{t}_{pq}$. Then by the definition of $\underline{t}_{pq}$ there must exist at least one branch excluding the feedback branch whose sending branch flow exceeds its branch capacity, which contradicts the fact that $\bar{f}_i \leq C_i$ for all $i \neq q$. Hence $\bar{f}_f = x_q/x_{max} \leq \underline{t}_{pq}$. Similarly the second part of Eq. (16) will be proved. Eq. (17) follows simply from $\underline{f}_f = \alpha_f \bar{f}_f$. Q.E.D.

If the number of branches in the shortest path from source p to sink q is found, better upper bounds on $\underline{t}_{pq}$ and $\bar{t}_{pq}$ may be obtained by the following theorem.

Theorem 6: Let n be the number of branches in the shortest path from source p to sink q. Let x'_n be the n^{th} smallest vertex potential in the same set as in Theorem 5. Then

$$\underline{t}_{pq} \leq x_q/x'_n \tag{18}$$

and

$$\bar{t}_{pq} \leq \alpha_f x_q/x'_n \; . \tag{19}$$

Proof: If all vertex potentials are normalized by x_n', there exists at least one branch in each path from p to q which has an initial vertex with a normalized potential greater than or equal to one, i.e., there is at least one branch in each path whose sending flow is greater than or equal to its branch capacity. This means that in every possible path from p to q, flow exceeds or equals its path capacity under the normalization by x_n'. Hence the theorem.

If all the possible paths from source p to sink q are found, another better upper bounds on $\underline{t}_{pq}$ and $\overline{t}_{pq}$ may be obtained by the following.

Theorem 7: Let $x_{1max}, x_{2max}, \ldots, x_{\ell max}$ be maximum initial vertex potentials in all possible paths from p to q, $1, 2, \ldots, \ell$ respectively. Let $x_s = \text{Min}\{x_{1max}, x_{2max}, \ldots, x_{\ell max}\}$. Then

$$\underline{t}_{pq} \leq x_q / x_s \tag{20}$$

and

$$\overline{t}_{pq} \leq \alpha_f x_q / x_s \; . \tag{21}$$

Proof: If all vertex potentials are normalized by x_s there exists at least one branch in each path which has $\overline{f} \geq C$. The proof follows in the same manner as in proof of Theorem 6.

Example 3: Now a simple numerical example is given to illustrate the theorems. Suppose that a lossy communication net shown in Fig. 6 is given and that we want to investigate the maximum flows from vertex 1 to vertex 3. The first of the two branch weights in Fig. 6 is the branch capacity and the other is the branch flow efficiency.

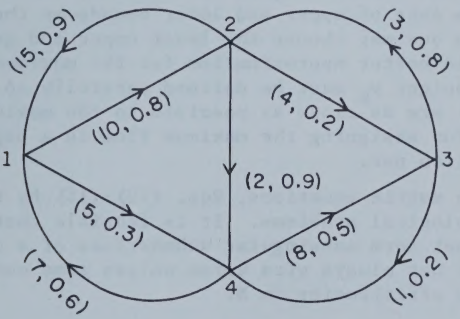


Fig. 6

Using the rules described earlier, the modified graph is obtained as shown in Fig. 1, where $y_1 = C_i$; $i = 1,2,3,4,5,f$, and numberical values are $C_1 = 10$, $C_2 = 4$, $C_3 = 8$, $C_4 = 5$, $C_5 = 2$, $C_f = 1$, $\alpha_1 = 0.8$, $\alpha_2 = 0.2$, $\alpha_3 = 0.5$, $\alpha_4 = 0.3$ and $\alpha_5 = 0.9$. The matrix equation corresponding to Eq. (12a) is

$$\begin{bmatrix} 15 & 0 & -\alpha_f & 0 \\ -8 & 6 & 0 & 0 \\ 0 & -0.8 & 1 & -4 \\ -1.5 & -1.8 & 0 & 8 \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} . \qquad (22)$$

α_f is determined from the condition for non-trivial solution, i.e., $\alpha_f = 4.97$. Hence $1/\alpha_f = 0.201$ is an equivalent efficiency of the net from vertex 1 to vertex 3 under the definition of branch weight y_k.

Solving Eq. (22) with $\alpha_f = 4.97$ for vertex potentials we have $x_1 = 4.97k$, $x_2 = 6.63k$, $x_3 = 15k$ and $x_4 = 2.424k$ where k is an arbitrary multiplicative constant. Since $x_q = x_3$, $x_{max} = x_2$, and $x_{min} = x_5$, Theorem 5 gives $2.26 \leq \underline{t}_{13} \leq 6.18$. Since $x_n' = x_s = x_1$ in this particular example, Theorems 3 and 4 give the same answer, $2.26 \leq \underline{t}_{13} \leq 3.02$ and $11.21 \leq \bar{t}_{13} \leq 15.0$. (The exact values are $\underline{t}_{13} = 2.45$ and $\bar{t}_{13} = 12.5$).

Note that if $y_2 = 6$ instead of 4 so that the maximum flow assignment will be proportional to the branch capacities at each vertex where more than one branches are directed away, then the lower bound in Theorem 5 gives the exact value of $\underline{t}_{13} = 2.85$ ($\bar{t}_{13} = 15$).

VI. CONCLUDING REMARKS

As was pointed out earlier, branch weight y_k can take any factor associated with branch k and for each defined y_k there are the four matrix equations Eqs. (12)-(15) for a lossy communication net. Suppose that m different branch weights are defined, then the 4m equations give the 4m sets of upper and lower bounds on the maximum flows in a net. Then one may choose the least upper and greatest lower bounds to obtain better approximation for the maximum flows. In this case branch weight y_k must be defined carefully so that the flows in Eqs. (4)-(7) are as close as possible to the maximum flow assignment. Method for assigning the maximum flow is a big problem in a lossy communication net.

Solutions of the matrix equations, Eqs. (12)-(15) by trees are also interesting topological problems. It is feasible that all the directed trees in a net form nonsingular submatrices of a connection matrix of the net but not always vice versa unless some conditions are placed on branch flow efficiencies in N.

ACKNOWLEDGMENT

The author is grateful for helpful discussions with Prof. W. Mayeda. The research reported herein was supported in part by AFOSR Grant AF-177-63 and in part by the Joint Services Program under Contract No. DA 28 043 AMC 00073(E).

REFERENCES

1. T. Fujisawa, "Maximal Flow in a Lossy Network," Proc. of the First Allerton Conference on Circuit and System Theory, Univ. of Illinois, Urbana, Illinois, September 1963.
2. W. Mayeda and M. E. Van Valkenburg, "Properties of Lossy Communication Nets," Report R-200, Coordinated Science Laboratory, University of Illinois, April, 1964, also to appear in the IEEE Trans. on Circuit Theory in 1965.
3. S. J. Mason, "Topological Analysis of Linear Nonreciprocal Networks," Proc. IRE, Vol. 45, pp. 829-838, June 1957.
4. F. E. Hohn, S. Seshu and D. D. Aufenkamp, "The Theory of Nets," IRE Trans. on Electronic Computers, Vol. EC-6, pp. 154-161,
5. C. L. Coates, "Flow Graph Solutions of Linear Algebraic Equations," IRE Trans. on Circuit Theory, Vol. CT-6, pp. 170-187, June 1959.
6. G. G. Dodd, "Principles and Applications of Unistor Graphs," Ph.D. Thesis, Dept. of Electrical Engineering, University of Illinois, 1965, also Coordinated Science Laboratory Report R-217, July 1964.
7. K. Onaga, "Analysis and Synthesis of Lossy Communication Nets," Ph.D. Thesis, Dept. of Electrical Engineering, University of Illinois, 1965, also Coordinated Science Laboratory Report R-245, March 1965.
8. W. S. Jewell, "Optimal Flow Through Networks with Gains," Operation Research, Vol. 10, No. 4, July-August 1962, pp. 476-499.

STOCHASTIC FLOW AND EFFICIENT TRANSMISSION THROUGH COMMUNICATION NETS*

KENJI ONAGA
Department of Electrical Engineering, University of Alabama
University, Alabama

ABSTRACT

Efficient transmission of stochastic flow of messages through store-and-forward type communication nets is disucssed in this paper. In order to express quality or performance of flow ψ quantitatively we define the penalty index $P(\psi)$ as the sum of total time delay $D(\psi)$, total transmission cost $K(\psi)$, and total transmission redundancy $R(\psi)$ with certain proportions of penalty. A necessary and sufficient condition of minimum penalty flow for the fixed flow value is obtained in Theorem 4 and is demonstrated by a simple example.

I. INTRODUCTION

As demonstrated in arrivals and services for telephone calls, telegraph messages, and customers at a supermarket, generation of information or service demands and time required for services are in general of stochastic nature. Stochastic flow is flow of such messages or service demands through nets. So far study of communication nets has been mainly focused on steady (or non-stochastic) flow. In the study of stochastic systems, average properties are the first thing to be concerned. Although by considering average flow, we can reduce properties of stochastic (or unsteady) flow to these of steady flow, there are many differences at a closer look. As a concrete example we take the following digital transmission of information through store-and-forward type communication nets.

II. THE MODEL: ASSUMPTIONS AND NOTATIONS

In this section we set up a mathematical model of the communication net by furnishing assumptions quantitatively and notations explicitly so that we can work on the model henceforth.

1. The message originates from the information source in a coded form and is fed to the net for transmission to the destination or sink. From the source a message occurs with constant probability in a unit time interval. For any time t the probability that a message occurs in

---

*This work was supported in part by the Joint Services Electronics Programs (U.S. Navy, and U.S. Air Force) under Contract NO-DA 28 043 AMC 0073(E). Portions of this work were also supported by Air Force Office of Scientific Research AF 49(638) - 1383. This work was done at the University of Illinois.

the small interval $(t, t + \Delta t)$ is $\lambda \cdot \Delta t + O(\Delta t)$, where λ is a constant and $O(\Delta t)$ is a quantity of smaller order of magnitude than Δt, and the probability that more than one message occurs is also smaller order of magnitude than Δt. Then it can be shown that the distribution function $A(t)$ of inter arrival time t of messages is given by $A(t) = 1 - e^{-\lambda t}$.

The average number of messages per second originating from the source is λ. This input is called Poisson distributed or a Poisson input because the number of messages in any time interval T has a poisson distribution with mean λT.

2. The length n of the message is governed by an exponential distribution function $B(n) = 1 - e^{-\mu n}$ where $\frac{1}{\mu}$ is the average length of the messages in bits. Throughout this paper we assume that the message length and the inter arrival time have no influence upon each other, in other words two distribution functions $A(t)$ and $B(n)$ are statistically independent.

3. At each station each outgoing channel has its own unlimited storage for incoming messages and the messages form a waiting line, called a queue. Messages in a queue are processed for transmission according to the first-come first-served principle.

4. Channels are assumed to be reliable and their capacities are measured by bits per second. Noise in a channel causes errors in the received message with probability q. At each station a message is received in a form of perforated tapes, is checked for possible errors, and is joined to a proper queue for retransmission if there is no error. Whenever error in a message is detected, retransmission of the message is asked to the immediate sending station until the correct message is received. Since average number of $\frac{1}{1 - q}$ $(= 1 + q + q^2 + \ldots)$ transmission is required in order to receive a single true message through the channel, for reception of ψ distinct messages total number (counting duplications) of $\psi/(1 - q)$ messages must be sent from the sending station, in other words the efficiency of the channel is $1 - q$ and the redundancy r is $q/(1 - q)$.

5. We assign two additional weights on each channel namely transmission cost κ cents per one message transmission and time delay d seconds per message. The transmission cost κ of a channel is constant independent of flow value of the channel. The time delay d is a monotone increasing function of flow value of the channel as to be shown later.

Natural choice of our model is a weighted linear graph where a station is represented by a node, a channel is represented by an edge, and the weights on the edge are capacity c, transmission cost κ and error probability q. If λ distinct messages enter a station for transmission through a channel with error probability q, then $\lambda/(1 - q)$ messages (counting duplications) leave the station. This seeming violation of flow conservation can be avoided if we stick to flow of distinct messages and reduce the edge capacity from c to $(1 - q)\cdot c$ in order to satisfy capacity limitation of flow. $\psi(e)$ of an average flow ψ or simply flow ψ is average number of distinct messages per second assigned by ψ on edge e.

III. TIME DELAY

Because of unsteadiness of flow it occurs with a positive probability that the number of incoming messages in a time interval exceeds the maximum number of messages the channel can handle in the time interval. As a consequent a message spends certain time in average in the queue waiting for transmission. The queueing theory provides numerous important results of steady states solutions of a simple queueing process. We are particularly interested in the average time spent in the queue and the transmitter, simply time delay per message d. It is given as follows [8] :

$$d = \frac{1}{(1 - q)\mu c - \lambda} = \frac{1}{\bar{c} - \lambda}$$

where λ is average input (distinct messages per second), $(1 - q)\mu c$ is the edge capacity measured by messages per second, and $\bar{c} = (1 - q)\mu c$.

As seen above the average time delay d per message of a channel with error probability q is a monotone increasing function of flow value and becomes infinite if $\lambda = \bar{c} = (1 - q)\mu c$. If the maximum allowable time delay for a channel is fixed to a certain value d_o, then we have $\lambda_o = \bar{c} - 1/d_o$. λ_o is the maximum number of messages per second the channel can handle within the fixed time delay d_o and is called effective edge capacity.

In general communication nets channels are connected each other in many ways and mathematical description of queueing process in such a multi-channel system are immensely complicated and evades our analysis. Main difficulty arises from lack of statistical independence between message inter arrival time and message length at internal (relay) stations. It is recognized that the source of dependency of the said two statistics lies in the assignment of a permanent length to each message during entire transmission process. Kleinrock [5] studied the effect of the independent assumption by the method of digital computer simulation and showed that for several elementary net configurations the introduction of independent assumption results in a rather insignificant change in the average message delay. It is proved by Burke [1] that the steady state output of a queue with Poisson inputs and exponential message length is itself Poisson distributed. Burke's Theorem coupled with the independent assumption enables us to consider each channel separately.

Suppose that flow ψ (in more precise, average flow ψ) has been assigned to the net N. The sending flow value is the average number of messages per second sent from source and the receiving flow value is the average number of messages per second received at sink. For there is no loss of flow in the net, these two values are of course equal, hence called the flow value λ of flow ψ. $D(\psi)$ of flow ψ is the time delay of ψ and considered as a measure of flow quality. $D(\psi)$ is defined by

$$D(\psi) = \sum_{e \varepsilon G} d_e \cdot \psi(e) \quad \text{where G is the edge set of N.}$$

Before undertaking a criterion of time delay quality, we need the following theorem.

Theorem 1: Let ψ_o and ψ' **be** flows with the same flow value λ, then it is possible to transform ψ_o into ψ' by successive applications of flow circulations. A flow circulation Δ is an assignment of flow increment along a oriented circuit of the net such that $\psi_o + \Delta$ is a flow. Such oriented circuit is called a flow circuit.

Proof: Define the difference $\phi_o = \psi_o - \psi'$. For convenience we call an edge zero, negative or positive depending on $\phi_o(e)$ is zero, negative or positive. Let N_1 be a net which is formed by removing zero edges from N. Start from any non-isolated node and trace negative edges in forward directions or positive edges in backward direction and stop if a node already traced is reached. We get an oriented circuit L_1. If $\psi_o \neq \psi'$, existence of the oriented circuit L_1 is guaranteed because flow conservation at all nodes including sink and source is satisfied by difference ϕ_o. We can assign flow circulation Δ_1, to L_1 such that $\psi_o + \Delta_1 = \psi_1$ is a flow and $\phi_o + \Delta_1 = \phi_1$ gives a zero label to an edge of L_1 while preserving the signs of the rest. Flow circulation Δ_1 decreases absolute values of ϕ_o of the edges of L_1. Construct N_2 by deleting zero edges of ϕ_1 from N_1 and trace a flow circuit L_2. Assign flow circulation Δ_2 along L_2 such that $\phi_1 + \Delta_2 = \phi_2$ gives a zero label to an edge of L_2 while preserving the signs of the rest. Now $\psi_2 = \psi_1 + \Delta_2 = \psi_o + \Delta_1 + \Delta_2$ is a flow. Repeat the procedure until N_{m+1} becomes empty. Therefore we have $\psi' = \psi_o + \Delta_1 + \Delta_2 + \ldots \Delta_m$.

We continue to use the same notations in the following discussion. Let ψ_o be given flow with flow value λ and let Δ_1 be a flow circulation of flow value ε_1 along a flow circuit L_1. The contribution of L_1 to the time delay $D(\psi_o)$ of ψ_o is

$$D_{L_1}(\psi_o) = \sum_{e \varepsilon L_1} d_{eo} \cdot \psi_o(e) = \sum_{e \varepsilon L_1} \frac{\psi_o(e)}{\bar{c}_e - \psi_o(e)}$$

where d_{eo} is the time delay of edge e evaluated by $\psi_o(e)$. The time delay $D(\psi_1)$ of flow $\psi_1 = \psi_o + \Delta_1$ is given by

$$D(\psi_1) = D(\psi_o) + \left\{ D_{L_1}(\psi_1) - D_{L_1}(\psi_o) \right\}$$

Let F_1 and B_1 be sets of forward edges and backward edges of L_1, respectively. Here an edge in L_1 is called forward (or backward) with respect to L_1 if the edge direction and the orientation of L_1 agree (or do not agree). Then, we have

$$D_{L_1}(\psi_1) - D_{L_1}(\psi_o) = \sum_{e \varepsilon F_1} \left\{ \frac{\psi_o(e) + \varepsilon_1}{\bar{c}_e - \psi_o(e) - \varepsilon_1} - \frac{\psi_o(e)}{\bar{c}_e - \psi_o(e)} \right\}$$

$$+ \sum_{e \varepsilon B_1} \left\{ \frac{\psi_o(e) - \varepsilon_1}{\bar{c}_e - \psi_o(e) + \varepsilon_1} - \frac{\psi_o(e)}{\bar{c}_e - \psi_o(e)} \right\}$$

$$= \sum_{e \varepsilon F_1} \bar{c}_e \cdot d_{eo} \cdot d_{e1} - \sum_{e \varepsilon B_1} \bar{c}_e \cdot d_{eo} \cdot d_{e1} \cdot \varepsilon_1$$

where $\bar{c}_e = (1 - q_e)\mu c_e$ and d_{ei} is the average time delay of edge e evaluated by $\psi_i(e)$.

Suppose that flow ψ' has the time delay $D(\psi')$ less than $D(\psi_o)$ and also that $\psi' = \psi_m$, $\psi_k = \psi_o + \sum_{i=1}^{k} \Delta_i$ where Δ_i is flow circulation of ε_i along a flow circuit L_i. Then we have

$$D(\psi') = D(\psi_o) + \left\{ D_{L_1}(\psi_1) - D_{L_1}(\psi_o) \right\} + \left\{ D_{L_2}(\psi_2) - D_{L_2}(\psi_1) \right\} + \ldots$$

$$+ \left\{ D_{L_m}(\psi_m) - D_{L_m}(\psi_{m-1}) \right\}$$

Since $D(\psi') < D(\psi_o)$, we must have for some j ($\leq$ m)

$$D_{L_j}(\psi_j) - D_{L_j}(\psi_{j-1}) = \varepsilon_j \left\{ \sum_{e \varepsilon F_j} \bar{c}_e \cdot d_{ej-1} \cdot d_{ej} - \sum_{e \varepsilon B_j} \bar{c}_e \cdot d_{ej-1} \cdot d_{ej} \right\} < 0$$

where F_j and B_j are sets of forward edges and backward edges of L_j, respectively. We know that F_j (or B_j) is composed of negative (or positive) edges of ϕ_{j-1} in N_j and that a negative (or positive) edge of ϕ_{j-1} is also a negative (or positive) edge of ϕ_o because of no sign change in the process of flow circulations. Now it is obvious that L_j is a flow circuit of the original flow ψ_o and that through the enitre process of flow circulations flow value of an edge in F_i $i = 1, 2, \ldots,$ m are increased successively and flow value of an edge in B_i $i = 1, 2, \ldots,$ m are decreased successively. Hence we have

$$d_{ei-1} \leq d_{ei} \quad \text{if} \quad e \varepsilon F_j \quad \text{and } 0 \leq i \leq j$$

$$d_{ei-1} \geq d_{ei} \quad \text{if} \quad e \varepsilon B_j \quad \text{and } 0 \leq i \leq j$$

Therefore for a flow circuit L_j of flow ψ_o we must have

$$d_{L_j}(\psi_o) = \sum_{e \varepsilon F_j} \bar{c}_e \cdot d_{eo}^2 - \sum_{e \varepsilon B_j} \bar{c}_e \cdot d_{eo}^2 < 0$$

and have proved the sufficiency of the following theorem.

Theorem 2. For the fixed flow value λ, flow ψ has the minimum time delay $D(\psi)$ if and only if there is no flow circuit L which has

$$d_L(\psi) = \sum_{e \varepsilon F} \bar{c}_e \cdot d_e^2 - \sum_{e \varepsilon B} \bar{c}_e \cdot d_e^2 < 0$$

where d_e is time delay per message of edge e evaluated by $\psi(e)$.

Proof: Necessity: If a flow circuit L of flow ψ has $d_L(\psi) < 0$, then we can choose a small positive number ε such that flow circulation Δ of ε along L produces a new flow $\psi' = \psi + \Delta$ which has time delay $D(\psi')$ less than $D(\psi)$. In other words under $d_L(\psi) < 0$ we can choose ε such that

$$D_L(\psi') - D_L(\psi) = \varepsilon \cdot \left\{ \sum_{e \varepsilon F} \bar{c}_e \cdot d_e \cdot d'_e - \sum_{e \varepsilon B} \bar{c}_e \cdot d_e \cdot d'_e \right\} < 0$$

where d'_e is evaluated by $\psi'(e)$.

Since ψ and ψ' have the same flow value, the above conclusion contradicts the minimality assumption of $D(\psi)$. Hence $d_L(\psi)$ must not be negative. Q.E.D.

IV. COST AND REDUNDANCY

Cost $K(\psi)$ and redundancy $R(\psi)$ of flow ψ are also measures of performance of flow and defined by

$$K(\psi) = \sum_{e \varepsilon G} k_e \cdot \psi(e)$$

$$R(\psi) = \sum_{e \varepsilon G} r_e \cdot \psi(e)$$

where k_e and $r_e = q_e/(1 - q_e)$ are also the transmission cost and redundancy of edge e and G is the edge set of the net.

We should notice that $K(\psi)$ and $R(\psi)$ behave exactly same way because they are defined in the same way and their increment due to flow circulation of ε along flow circuit L are given similarly by

$$k_L(\psi) \cdot \varepsilon = k_L \cdot \varepsilon = \varepsilon \cdot \left\{ \sum_{e \varepsilon F} k_e - \sum_{e \varepsilon B} k_e \right\}$$

$$r_L(\psi) \cdot \varepsilon = r_L \cdot \varepsilon = \varepsilon \cdot \left\{ \sum_{e \varepsilon F} r_e - \sum_{e \varepsilon B} r_e \right\}$$

Minimization of $K(\psi)$ or $R(\psi)$ for the fixed flow value λ has been known as the minimum transportation cost problem [2] and many characterizations have been stated in different forms. For our purpose the following form is convenient and can be proved easily by the same line of reasonings as for Theorem 2.

Theorem 3: For the fixed flow value λ, flow ψ has minimum $K(\psi)$ or $R(\psi)$ if and only if no flow circuit L has $k_L < 0$ or $r_L < 0$, respectively.

V. MINIMIZATION OF THE PENALTY INDEX

Efficient transmission of message flow must take time delay $D(\psi)$ transmission cost $K(\psi)$ and transmission redundancy $R(\psi)$ into consideration with certain proportions of penalty. The penalty index $P(\psi)$ of flow ψ is defined by $P(\psi) = \alpha \cdot D(\psi) + \beta \cdot K(\psi) + \partial \cdot R(\psi)$, $\alpha + \beta + \partial = 1$ and $\alpha, \beta, \partial \geq 0$. The increment $p_L(\psi) \cdot \varepsilon$ of $P(\psi)$ due to a flow circulation of ε along a flow circuit L is expressed by

$$p_L(\psi) \cdot \varepsilon = \{\alpha \cdot d_L(\psi) + \beta \cdot k_L + \partial \cdot r_L\} \cdot \varepsilon$$

$$= \varepsilon \cdot \left\{ \sum_{e \varepsilon F} (\alpha \cdot \bar{c}_e \cdot d_e^2 + \beta \cdot k_e + \partial \cdot r_e) - \sum_{e \varepsilon B} (\alpha \cdot \bar{c}_e \cdot d_e^2 + \beta \cdot k_e + \partial \cdot r_e) \right\}$$

Now the following theorem is obvious.

Theorem 4: For the fixed flow value λ, flow ψ has the minimum penalty index if and only if there is no flow circuit L whose penalty $p_L(\psi)$ is negative.

It should be noted that the penalty index $P(\psi)$ of flow ψ is a nonlinear function of ψ and that $P(\psi)$ is a generalization of linear transmission (or transportation) cost. Theorem 4 provides a good criterion of minimum penalty for a given flow but does not provide a flow assignment algorithm which successively yields flows with minimum penalty index. For this purpose the following theorem is important.

Theorem 5: Suppose that flow ψ with flow value λ and minimum penalty index has been already assigned to the net. If Γ is a unique flow path from source to sink with the minimum penalty $p_{\Gamma}(\psi)$ and if Δ is a flow increment of a small flow value ε along Γ, then flow $\psi + \Delta$ has the minimum penalty index for the flow value $\lambda + \varepsilon$. Here $p_{\Gamma}(\psi)$ is defined exactly same as $p_L(\psi)$ above.

In the case of linear penalty, an adjective "unique" can be omitted from the statement and the theorem leads us to a dynamic program of flows with minimum penalty. The proof can be furnished by either the primal - dual algorithm of linear programming [2] [3] [4] or a pure graph theoretic approach [7] . In our case, the adjective "unique" is necessary and the proof is the same as above.

VI. AN ILLUSTRATIVE EXAMPLE

Let us illustrate the theorems obtained above by a simple example shown in Fig. 1. For each edge capacity (bits per second), error probability, and transmission cost (cents per message) are given in the above order by e_1: (40, 0.1, 2), e_2: (60, 0.2, 1), e_3: (60, 0.1, 3), e_4: (40, 0.3, 3), and e_5: (30, 0.2, 1). If we take the maximum allowable time delay d_o through the edge as 1 second per message and μ as 0.1, we can calculate effective edge capacity $\bar{c}_i$ from the equation $\bar{c}_i = \bar{e}_{e_i} - 1/d_o = (1 - q_{e_i})\mu\, c_{e_i} - 1/d_o$: $\bar{c}_1 = 2.6$, $\bar{c}_{e_1} = 3.6$; $\bar{c}_2 = 3.8$, $\bar{c}_{e_2} = 4.8$; $\bar{c}_3 = 4.4$, $\bar{c}_{e_3} = 5.4$; $\bar{c}_4 = 1.8$, $\bar{c}_{e_4} = 2.8$; $\bar{c}_5 = 1.4$, $\bar{c}_{e_5} = 2.4$; (messages per second). For a given flow ψ in Fig. 2, the penalty index $P(\psi)$ is calculated as follows with the choice of $\alpha = 0.1$, $\beta = 0.2$, and $\partial = 0.7$. $P(\psi) = 0.1 \times 8.65 + 0.2 \times 24.6 + 0.7 \times 2.55 = 7.552$. An oriented circuit L in Fig. 2 is the only flow circuit in the net and its penalty $p_L(\psi)$ is positive as below, hence by Theorem 4 the given flow ψ has the minimum penalty for flow value 5.6 (messages per second) which is the maximum value attainable.

$$p_L(\psi) = 0.675$$

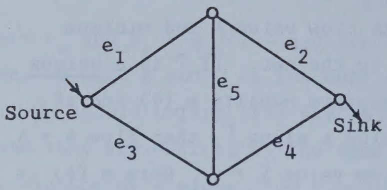


Fig. 1. An illustrative example

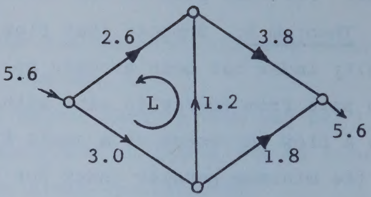


Fig. 2. A given flow. Solid arrows indicate saturation.

VII. CONCLUSIONS

In this paper we have considered a particular mode of operation of communication nets namely restoration scheme of error messages by means of retransmissions. Efficient transmission of flow ψ can be achieved by avoiding flow circuits L with negative penalty $p_L(\psi)$, Theorem 4. Among other possible modes of operations, the following two modes are important: mode A is a scheme that at the receiving stations error messages are dismissed from the queues without any consequence and mode B is a scheme that at the receiving stations no request of retransmission and no dismissal of error messages take place. It must be noticed that in either mode of operation flow of correct messages suffers loss during process of transmission and obeys the axioms of flow in lossy communication nets [6], [7]. Theorem 1 is no longer valid in these cases and there are considerable alterations and difficulties in the formulation of corresponding theorems cited in this paper.

ACKNOWLEDGEMENT

The author is indebted to Professors W. Mayeda and M. E. Van Valkenburg, University of Illinois, for many helpful suggestions.

REFERENCES

1. Burke, P. J., The Output of a Queueing System, J. Opns. Res. Soc. Amer., vol. 4, pp. 699-704, 1956.
2. Ford, J. R. and Fulkerson, D. R., Flows in Networks, New Jersey, Princeton University Press, 1962.
3. Iri, M., A New Method of Solving Transportation - Network Problems, J. Op. Res. Soc. Japan. vol. 3, 1960, pp. 27-87.

4. Jewell, W. S., Optimal Flow Through Networks, Interim Technical Report No. 8, M.I.T., 1958.

5. Kleinrock, L., Communication Nets, New York, McGraw-Hill, 1964.

6. Mayeda, W. and Van Valkenburg, M. E., Properties of Lossy Communication Nets, To appear in IEEE Transaction of Circuit Theory.

7. Onaga, K., Dynamic Programming of Optimum flows in Lossy Communication Nets, Report R-247, Coordinated Science Lab., University of Illinois, Urbana, Illinois, March, 1965.

8. Riordan, J., Stochastic Service Systems, New York, John Wiley and Sons, 1962.

TRAFFIC NETWORK ANALYSIS

J. L. GOODNUFF
Department of Electrical Engineering
Michigan State University
East Lansing, Michigan

W. B. HALL
Crouse-Hinds Company

ABSTRACT

Electrical engineers have developed analytical tools and procedures for the analysis of electric networks, only to find that they are also useful in the analysis of many other types of systems. Electric network techniques are now universally applied in the analysis of mechanical, hydraulic, thermal, areonautical and nearly all other types of physical systems.

It has recently become apparent that those same analysis techniques can be applied to certain types of socio-economic systems. This paper demonstrates the application of these techniques in establishing state models characterizing the dynamics traffic networks, which may serve as a basis for design and simulation studies.

It is hoped that this paper will serve to broaden the perspective of network engineers and encourage more investigations into the applicability of existing network techniques in the analysis of non-physical systems.

ITERATIVE CHEBYSHEV APPROXIMATION TECHNIQUE FOR EQUALIZER SYNTHESIS*

G.C. TEMES and J.A.C. BINGHAM
Stanford Linear Accelerator Center, Stanford University
Stanford, California

Abstract: The Remez-Werner algorithm is extended to the Chebyshev approximation of general loss- and phase-characteristics. The resulting technique is useful for the design of equalizers, pulse shaping networks, impedances and delay lines.

The application of the process to the synthesis of a video cable system is described.

I. INTRODUCTION

A common problem in network synthesis is the approximation of a numerically or explicitly known function in terms of a realizable network characteristic, such as loss, phase, voltage ratio, driving-point immittance, etc. The network characteristic must be either an n^{th} degree rational function $P_{n-m}(\omega)/Q_m(\omega)$ of ω, or a simple function (log, $\tan^{-1}$ etc.) of P_{n-m}/Q_m .

The main purpose of this paper is to show how an arbitrary function $f(\omega)$, continuous in the frequency-range of interest $[\omega_A,\omega_B]$, can be approximated with a network function $R(\omega)$ in a weighted Chebyshev manner, i.e., so that

$$\max_{\omega_A \leq \omega \leq \omega_B} \left| w(f-R) \right| = \max_{[\omega_A,\omega_b]} \left| \Delta \right| = \min \tag{1}$$

In (1), $w(\omega)$ is the weight and Δ is the weighted error. It can be shown[1-4] that for given w and f, Δ (and hence R) can be obtained by the following procedure:

α. select Δ^o with $n+1$ zeros in $[x_A,x_B]$;

β. select $n+2$ points $[x_j^1]$ in $[x_A,x_B]$ on which Δ^o alternates; include the x_k^1 for which $\left|\Delta^o(x_k^1)\right| = \max\limits_{[x_A,x_B]} \left|\Delta^o\right|$;

γ. solve

$$\Delta^1 \, (x_j^1) = (-1)^j \, E^1 \qquad j = 1, 2, \ldots, n+2 \tag{2}$$

for Δ^1 , E^1 ;

δ. find, from $\Delta^1(x)$, a new set $[x_j^2]$ such that the x_k^2 satisfying $\Delta^1(x_k^2)| = \max\limits_{[x_A,x_B]} \left|\Delta^1\right|$ is included, and also such that $\min\limits_j \left|\Delta^1(x_j^2)\right| \geq E^1$;**

---

[*] Work supported by the U. S. Atomic Energy Commission.

[**] This can be done by replacing one of $[x_j^1]$ by x_k^2; or by selecting the $n+2$ local maxima of $|\Delta^1|$. The latter process usually gives faster convergence.

ε. repeat β, γ, δ, until $[x_j^{\ell}] = [x_j^{\ell+1}]$, and/or $E^{\ell} = E^{\ell+1}$

with a prescribed accuracy.

This process is called the second Remez algorithm. [2] For network applications, usually $x=\omega^2$. The convergence conditions are well-known [1-4]; essentially, the continuity of f, w and all approximants is required in the approximation range $[x_A, x_B]$.

The discussion will be restricted to methods for carrying out the Remez procedure, involving the solution of linear and single-variable polynomial equations only ("linear" methods).

II. DIRECT METHODS FOR SOLVING EQ. (2)

The most difficult part of the Remez algorithm is step γ. For network functions this is equivalent, in general, to solving

$$w(x_j)\left[f(x_j) - F\left(\frac{P_{n-m}(x_j)}{Q_m(x_j)}\right)\right] = (-1)^j E;\ j = 1,2,\ldots n+2 \qquad (3)$$

for the coefficients a_i, b_i of P_{n-m} and Q_m, as well as for E. From (3)

$$P_{n-m}(x_j) - Q_m(x_j)\, F^{-1}\left[f(x_j) - (-1)^j \frac{E}{w(x_j)}\right] = 0;\ j = 1,2,\ldots n+2 \qquad (4)$$

In terms of the matrices, X_k, a, b and G which are defined by

$$X_k = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n+2} & x_{n+2}^2 & \ldots & x_{n+2}^k \end{pmatrix};\quad a = \begin{pmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_{n-m} \end{pmatrix};\quad b = \begin{pmatrix} 1 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{pmatrix};$$

$$G = \operatorname{diag}(g_1, g_2, \ldots g_{n+2});\quad g_j = F^{-1}\left[f(x_j) - (-1)^j \frac{E}{w(x_j)}\right],$$

eq. (4) can be written in the form

$$X_{n-m}\, a - G X_m b = 0 \qquad (5)$$

Eq. (5) represents a set of linear equations in a and b alone; however, in general (5) is nonlinear in E. The straightforward method for finding E is to treat (5) as an overdetermined set of n+2 linear equations for the n+1 unknown elements of a and b. For consistency, then, the determinant equation

$$D = \left| X_{n-m} \,\middle|\, G X_m \right| = 0 \qquad (6)$$

must hold. In (6), the only unknown is E; it can (in principle) be obtained from (6), and after that the linear system (5) can be solved for a and b.

D can be expanded directly, using Laplace and Vandermonde expansions, in the following steps:

(1) Select $m+1$ integers from $S = [1,2, \ldots, n+2]$. Call their set S_{m+1} and the set of the remaining $n-m$ integers, $\bar{S}_{m+1}$, i.e., $S - S_{m+1} = \bar{S}_{m+1}$.

(2) Find $\delta = \left[\prod_{j \in S_{m+1}} g_j \right] \left[\prod_{\substack{j,k \in S_{m+1} \\ j>k}} (x_j - x_k) \right] \left[\prod_{\substack{j,k \in \bar{S}_{m+1} \\ j>k}} (x_j - x_k) \right]$

(3) Then $D = \sum_{\text{All } S_{m+1}} (-1)^{\Sigma_{j \in S_{m+1}} j} \delta$.

Equation (6) usually leads to an (m+1) degree polynomial equation for E (or for a simple function of E). This equation can sometimes be expediently obtained in the following steps:

(1) Select an $i \in S$; calculate E_i such that $g_i(E_i) = g_{ii} = 0$.

(2) Calculate $g_{ji} = g_j(E_i)$.

(3) Select S_{m+1} not including i ; find $\delta(E)$ and $D(E_i) = D_i$ as before.

(4) Calculate D_i for $m+2$ values of i ; find the $m+1$ degree interpolating polynomial from these $m+2$ function-values.

III. WERNER'S ALGORITHM [5]

In practice, the direct methods described in Section II are tedious; also, they generally lead to a set E values, all satifying (6). Due to uniqueness,[1] at most one of these can lead to a continuous approximant. Although for accurate approximations usually the E with the lowest absolute value is the useful one, there are exceptions.

For the above reasons, an alternative method* is frequently preferable. Matrix L

$$L = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n+2}) \tag{7}$$

is found such that

$$X_m^T L X_{n-m} = 0 \tag{8}$$

(0 is the null-matrix). Equation (8) can be rewritten as

$$X_n^T \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n+2} \end{pmatrix} = 0 \tag{9}$$

---

*Developed by H. Werner [5] for direct rational approximation. Other iterative methods have also been used. [6,7]

Equation (9) can be solved for the λ_i using Vandermonde's expansion [*]

$$\lambda_i = \prod_{\substack{j=1 \\ j\neq i}}^{n+2} (x_i - x_j)^{-1} \tag{10}$$

Multiplying (5) by $X_m^T L$ and using (8)

$$X_m^T LGX_m b = Ab = 0 \tag{11}$$

A is an $(m+1) \times (m+1)$ matrix; hence, (11) is a system of $m+1$ equations for b and E . As with (6), using

$$|A| = 0 \tag{12}$$

for finding E leads to difficulties. However, the scalar equation

$$b^T X_m^T LGX_m b = \sum_{j=1}^{n+2} \lambda_j g_j [Q_m(x_j)]^2 = 0 \tag{13}$$

can usually easily be solved for E , if b is known. Even if only an approximately correct b is given (in practice, the one obtained in the preceding iteration), the error in E is normally small.

Having obtained E , (11) can be used to find an improved b , either by discarding one (inside) equation[**], or, preferably, by solving the full set (11) for least square error. From (5), a can be now obtained, either by discarding m+1 equations[***] or, preferably, using the least squares approach. Next, the new $[x_j]$ can be found from a and b , a new E calculated from (13) and the process repeated.

IV. REDUCTION TO RATIONAL APPROXIMATION

It will be shown in Section V that, for rational approximating function, the application of the Remez algorithm is relatively straightforward. Hence, it is useful to develop a procedure to reduce the general problem (3) to that of a rational approximation. Assuming that the equal-ripple error $\Delta^*(x)$ is achieved,

$$w(x)\left[f(x) - F\left(\frac{P_{n-m}(x)}{Q_m(x)}\right)\right] = \Delta^*(x) \tag{14}$$

If, for $x_A \leq x \leq x_B$,

$$\left|\frac{\Delta^*(x)}{w(x)}\right| \ll \left|\frac{dF^{-1}(f)}{df} \bigg/ \frac{d^2F^{-1}(f)}{df^2}\right| \tag{15}$$

---

[*] Naturally, any proportionality factor can be associated with the $[\lambda_i]$.

[**] b contains only the m unknowns $b_1, \ldots b_m$.

[***] Evenly spread within the set.

then

$$F^{-1}\,[f(x)] \cong \frac{P_{n-m}}{Q_m} + \frac{\Delta^*(x)}{w(x)}\;\frac{dF^{-1}(f)}{df} \tag{16}$$

so that the rational approximation problem

$$w(x)\left[\frac{dF^{-1}(f)}{df}\right]^{-1}\left\{F^{-1}[f(x)] - \frac{P_{n-m}(x)}{Q_m(x)}\right\} = \Delta^*(x) \tag{17}$$

is equivalent to the original one given in (14) to a first-order approximation.

The inverse problem can also arise, when one wants to cancel a weighting factor in a rational approximation, by introducing F, i.e., given w' and h in

$$w'(x)\left[h(x) - \frac{P_{n-m}(x)}{Q_m(x)}\right] = \Delta^*(x) \tag{18}$$

F, f are to be found such that (18) is transformed into (14), with $w(x) \equiv 1$. From (17) - (18), then

$$f = F(h) = \int w'dh \tag{19}$$

V. APPROXIMATIONS USING NETWORK FUNCTIONS

A. Modulus of Transfer Function, Prescribed Loss Poles

This problem reduces to a polynomial approximation;[3] if $f(x)$ is the squared modulus of the transfer function (voltage or current ratio, operating factor, etc.), then

$$\Delta(x) = w(x)\left[f(x) - \frac{P_n(x)}{g(x)}\right] \tag{20}$$

Here (5) is linear in a and E. The function itself is obtained from P_n/g by Hurwitz factorization.

B. Loss, Prescribed Poles

Here f is the loss, and

$$\Delta(x) = w(x)\left\{f(x) - \log\left[\frac{P_n(x)}{g(x)}\right]\right\} \tag{21}$$

For (5), here

$$X_m\ b = \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix},\quad g_j = g(x_j)\ \exp\left[f(x_j) - (-1)^j\,\frac{E}{w(x_j)}\right] \tag{22}$$

So, by (13), if $w(x) \equiv 1$,

$$E = \frac{1}{2} \log \left\{ \frac{\sum\limits_{\text{all even } j} \left|\lambda_j\right| g(x_j) \exp[f(x_j)]}{\sum\limits_{\text{all odd } j} \left|\lambda_j\right| g(x_j) \exp[f(x_j)]} \right\} \tag{23}$$

If a weighting function is present, the method of Sec. IV can be used. From (17),

$$\Delta = w(x) \left[1 - \frac{P_n(x)}{g(x) \exp[f(x)]} \right] \tag{24}$$

describes the equivalent rational approximation problem.

C. Modulus of Transfer Function, Arbitrary Loss Poles

Now

$$\Delta(x) = w(x) \left[f(x) - \frac{P_{n-m}(x)}{Q_m(x)} \right] \tag{25}$$

and

$$g_j = f(x_j) - (-1)^j \frac{E}{w(x_j)} \tag{26}$$

(6) can be solved using the second direct method with $E_i = (-1)^i w(x_i) f(x_i)$. Using Werner's method, (11) turns out to be an eigenvalue equation of a real symmetric matrix, with E as the eigenvalue and b as the eigenvector. Hence, all $m+1$ E values are real. From (13) and (8),

$$E = \frac{b^T X_m^T L \overbrace{\left\{ \operatorname{diag}[f(x_j)] \right\}}^{F} X_m b}{b^T X_m^T L \underbrace{\left\{ \operatorname{diag}\left(\frac{(-1)^j}{w(x_j)} \right) \right\}}_{W} X_m b} = \frac{\sum\limits_{j=1}^{n+2} \lambda_j f(x_j) \left[Q_m(x_j)\right]^2}{\sum\limits_{j=1}^{n+2} \lambda_j \frac{(-1)^j}{w(x_j)} \left[Q_m(x_j)\right]^2} \tag{27}$$

or

$$E = \frac{b^T X_m^T L \left(F X_m b - X_{n-m} a\right)}{b^T X_m^T L W X_m b} = \frac{\sum\limits_{j=1}^{n+2} \left|\lambda_j\right| \left[Q_m(x_j)\right]^2 \left|\Delta(x_j)\right| / w(x_j)}{\sum\limits_{j=1}^{n+2} \left|\lambda_j\right| \left[Q_m(x_j)\right]^2 / w(x_j)} \tag{28}$$

Hence, E is the weighted average of the $\left|\Delta(x_j)\right|$.

D. Loss, Arbitrary Poles

For this case,

$$\Delta(x) = w(x) \left\{ f(x) - \log \left[\frac{P_{n-m}(x)}{Q_m(x)} \right] \right\} \tag{29}$$

and

$$g_j = \exp\left[f(x_j) - (-1)^j \frac{E}{w(x_j)} \right] . \tag{30}$$

Using the first direct method, (6) turns out to be an $m+1$ degree polynomial equation* in $\exp(2E)$, if $w(x) \equiv 1$. All roots are real and positive**; however, only one can lead to a continuous $[P_{n-m}(x)]/[Q_m(x)]$.

From (13), for $w(x) \equiv 1$,

$$E = \frac{1}{2} \log \left\{ \frac{\sum\limits_{\text{all even } j} |\lambda_j| \exp[f(x_j)][Q_m(x_j)]^2}{\sum\limits_{\text{all odd } j} |\lambda_j| \exp[f(x_j)][Q_m(x_j)]^2} \right\} . \tag{31}$$

If a weighting function has to be included, the reduction to a rational approximation is expedient. Eq. (17) now gives

$$\Delta(x) = w(x) \exp[-f(x)] \left\{ \exp[f(x)] - \frac{P_{n-m}(x)}{Q_m(x)} \right\} \tag{32}$$

as the error-function of the equivalent rational approximant. Conversely, the minimization of relative error with a rational approximant can be transformed into the unweighted loss-approximation, using (19).

E. Phase

For phase-approximation, with $\Phi \equiv f$

$$\Delta(x) = w(x) \left\{ f(x) - \tan^{-1} \left[\sqrt{x} \frac{P_{n-m}(x)}{Q_m(x)} \right] \right\} , \tag{33}$$

and

$$g_j = x_j^{-\frac{1}{2}} \frac{\tan[f(x_j)] - (-1)^j \tan\left[\frac{E}{w(x_j)}\right]}{1 + (-1)^j \tan[f(x_j)] \tan\left[\frac{E}{w(x_j)}\right]} \tag{34}$$

For $w(x) \equiv 1$, (6) gives an $n+2$ degree polynomial equation in $\tan E$. Choosing, for all $n+2$ i values, $E_i = (-1)^i f(x_i)$, the second direct method can be used to obtain this equation.

The Werner method is not very practical for this case, since (13) turns out to be a polynomial, rather than linear, equation for $\tan E$.

If $w(x) \not\equiv 1$, the reduction to rational approximation is advisable.

---

[*] If $m \leq (n+1)/2$, which is necessary for realizability. Otherwise the degree is the integer part of $(n+3)/2$, and $\exp(2E) = 0$ is a root.

[**] This can be seen from (13), if b^T is replaced by its complex conjugate.

From (17), now

$$\Delta(x) = w(x)\ \sqrt{x}\ \cos^2[\,f(x)]\left\{x^{-\frac{1}{2}}\ \tan[f(x)] - \frac{P_{n-m}(x)}{Q_m(x)}\right\}. \qquad (35)$$

Eq. (35) can be directly applied if $|f| < \frac{\Pi}{2}$; otherwise, special methods must be used.

For LC twoports, the resulting $P_{n-m}(x)$ and $Q_m(x)$ must have all their zeros on the positive real x-axis interlaced.* However, it is known that for linear phase the Chebyshev approximant is realizable; hence, adding sufficient linear phase ($k\omega$) to Φ will make the approximant of $f(x) + k\sqrt{x}$ also realizable.

F. Phase-Delay

Now

$$\Delta(x) = w(x)\left\{f(x) - x^{-\frac{1}{2}}\ \tan^{-1}\left[\sqrt{x}\ \frac{P_{n-m}(x)}{Q_m(x)}\right]\right\} \qquad (36)$$

and

$$g_j = x_j^{-\frac{1}{2}}\ \frac{\tan\left[\sqrt{x_j}\ f(x_j)\right] - (-1)^j\ \tan\left[\dfrac{\sqrt{x_j}\ E}{w(x_j)}\right]}{1 + (-1)^j\ \tan\left[\sqrt{x_j}\ f(x_j)\right]\tan\left[\dfrac{\sqrt{x_j}\ E}{w(x_j)}\right]}\ . \qquad (37)$$

Except for $w(x) = k\ \sqrt{x}$, (6) and (13) are now nonlinear in tan E. The only method that leads to a linear problem is that of Sec. IV. By (17), then

$$\Delta(x) = w(x)\ \cos^2[\sqrt{x}\ f(x)]\left\{x^{-\frac{1}{2}}\ \tan\ [\sqrt{x}\ f(x)] - \frac{P_{n-m}(x)}{Q_m(x)}\right\} \qquad (38)$$

gives the error of the equivalent rational approximation.

The remarks made in connection with (35) are here also applicable.

VI. APPLICATION TO THE EQUALIZATION OF COAXIAL CABLES

It is often required that an equalized cable transmit square pulses, but the general problem of designing an equalizer to mimimize rise time, overshoot and final value attenuation has not been solved. A simplistic approach is to merely define the desired attenuation as that which, together with the minimum associated phase, results in a step response with a rise time and overshoot less than some given amounts and then to adjust the loss to this.

It is known that if a low pass system is band-limited then an attenuation which is proportional to frequency squared (i.e., Gaussian) up to a loss of 20 db or more is in a sense optimum.[8] However, since a coaxial cable is only very gradually band-limited, it is not practical to make the loss Gaussian up to more than a few db; therefore, the shape of the attenuation curve beyond the band of fitting will affect the pulse response and some exponent of frequency other than two may be preferable for the desired attenuation within the band.

A simplified model of a coaxial cable which only takes account of losses due to the skin effect shows that the attenuation is proportional to $\sqrt{f}$.[9] Actual cables do not obey this relationship exactly but it is the only one for which the step response can be calculated simply and with confidence.

---

*These zeros are at or near the abscissae of the $\Phi = k\Phi$ and $\Phi = k\Pi + \frac{\Pi}{2}$ points, respectively.

An experiment was performed to find the best values of the exponent of frequency in the desired attenuation. A length of cable was "equalized" to losses proportional to $f^{1.8}$, $f^{2.0}$, $f^{2.4}$, $f^{3.0}$ and $f^{4.0}$, and the step response of cascades of one to five of these equalized systems* was computed. The loss of the cable (100 ft. of RG58C/U) was taken to be the idealized $6.0 \sqrt{f \times 10^{-8}}$ db and therefore the function of x**which was to be approximated, was $f(x) = \alpha x^{\gamma} - 6x^{.25}$ for $\gamma = .9, 1.0, 1.2, 1.5,$ and 2.0; α was chosen so that $f'(1) = 0$ to induce a smooth transition between the attenuation in the adjusted and unadjusted bands. The form of the desired attenuation in typical cases of $\gamma <, =,$ or > 1.0 is shown in Fig. 1; the form of the error function for an equalizer with $n = 6$ is shown in Fig. 2. The results of the syntheses and time domain analyses are summarized in Table I.

TABLE I

100 ft RG58 C/U Unequalized Attenuation $= 6.0 \sqrt{f \times 10^{-8}}$ db
Unequalized 10% - 90% r.t. = 45 nsec
Adjusted from .3 to 100 mcs so that

$$10 \log (P/Q) = A_0 + \alpha x^{\gamma} - 6x^{.25} + \Delta(x)$$

and

$$|\Delta(x)| \leq E$$

			← one equalized system →			← five systems →	
γ	α	Edb	10-90 r.t.	Overshoot	dc value	10-90 r.t.	Overshoot
0.9	1.66	0.026	2.3 nsec	0.2%	0.588	5.9 nsec	0.25%
1.0	1.5	0.028	2.2 nsec	0.36%	0.575	5.5 nsec	1.1%
1.2	1.25	0.032	2.0 nsec	0.7%	0.555	4.7 nsec	1.9%
1.5	1.0	0.036	1.8 nsec	1.0%	0.533	4.2 nsec	3.4%
2.0	0.75	0.040	1.6 nsec	1.7%	0.511	3.7 nsec	6.2%
∞	0	0.029	1.0 nsec	2.3%	0.393	2.2 nsec	7.4%

It can be seen from Table I that for an overshoot less than 1% and for one length of cable and one equalizer, the largest (i.e., "fastest") value of γ that can be used is 1.5, whereas for a system of five lengths of cable and five equalizers in tandem, the largest value is slightly less than 1.0 (or near Gaussian).

Experiments were performed for the cases of $\gamma = 1$ and $\gamma = \infty$. The circuits were built and tested; cascaded systems were simulated on the signal circulating setup of Fig. 3. The experimental results confirmed the calculations.

It is probable that some function more complicated than αx^{γ} would be better (i.e., resulting in less overshoot for the same rise time or vice versa) as the overall attenuation. However, the simple function chosen provides one parameter (γ) which can be varied easily to obtain a near optimum attenuation "equalizer" for a given system requirement.

REFERENCES

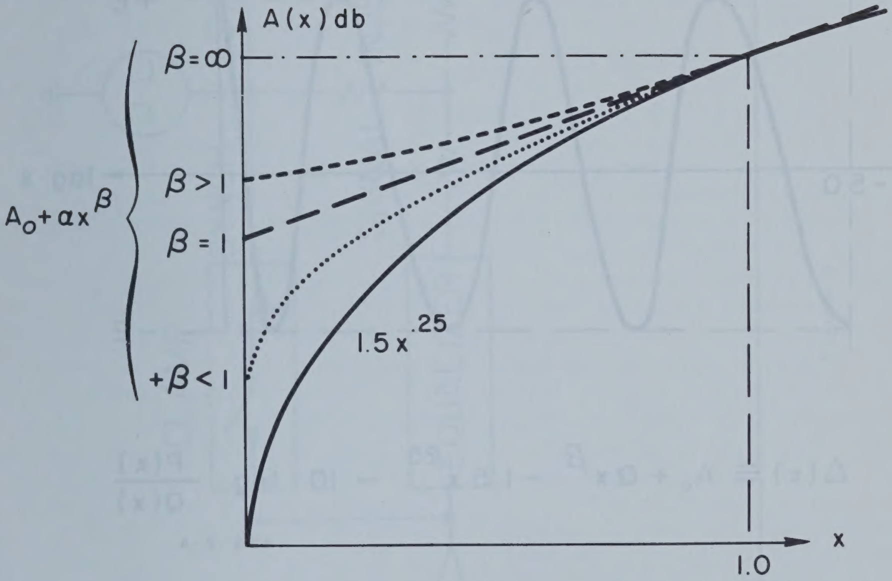
1. Novodvorskii, E.P., and Pinsker, I.S., "The process of equating maxima," Uspehi Mat. Nauk. 6 (1951), pp. 174-181 (English trans. by A. Shenitzer).
2. Remés, E., "Sur le calcul effectif des polynomes d'approximation de Tchebychef." C.R. Acad. Sci., Paris 199, pp. 337-340, 1934.

---

*This was done to magnify any overshoot in the step response.

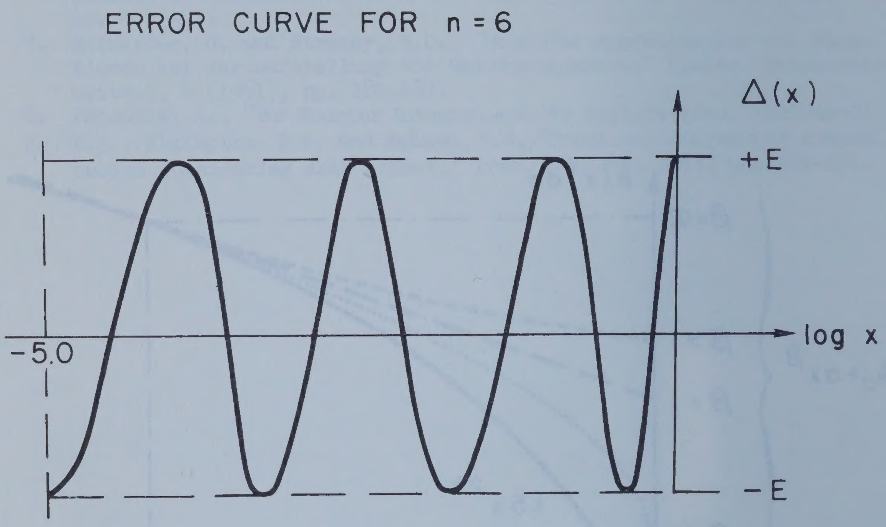
**$x = (\omega/2\pi \times 10^8)^2$.

3. Shenitzer, A., "Chebyshev approximation of a continuous function by a class of funtions." J. Assoc. Comput. Mach., 4, pp. 30-35.
4. Achieser, N.I., "Theory of Approximation," 1956, Frederick Ungar Publishing Co., New York.
5. Werner, H., "Rationale Tschebyscheff-Approximation, Eigenwerttheorie, und Differenzenrechnung."Arch. Rat. Mech. Anal., 13, 1963,pp. 330-347.
6. Fraser, W., and Hart, J.F., "On the computation of rational approximations to continuous functions." Comm. Ass. Comput. Mach., 5, pp. 401-403.
7. Meinardus, G. and Strauer, H.D., "Uber die Approximation von Funktionen bei der Aufstellung von Unterprogrammen," Elektr. Datenverarbeitung, 4 (1961), pp. 180-187.
8. Papoulis, A., "The Fourier Integral and its Applications," McGraw-Hill,p.104.
9. E.g., Wigington R.L. and Nahman, N.S.,"Transient analysis of coaxial cables considering skin effect," Proc. IRE, Feb. 1957, pp. 166-174.



FITTING OF DESIRED ATTENUATION TO CABLE ATTENUATION

Figure 1



$$\Delta(x) \equiv A_o + \alpha x^{\beta} - 1.5\, x^{.25} - 10 \log \frac{P(x)}{Q(x)}$$

424-2-A

Figure 2

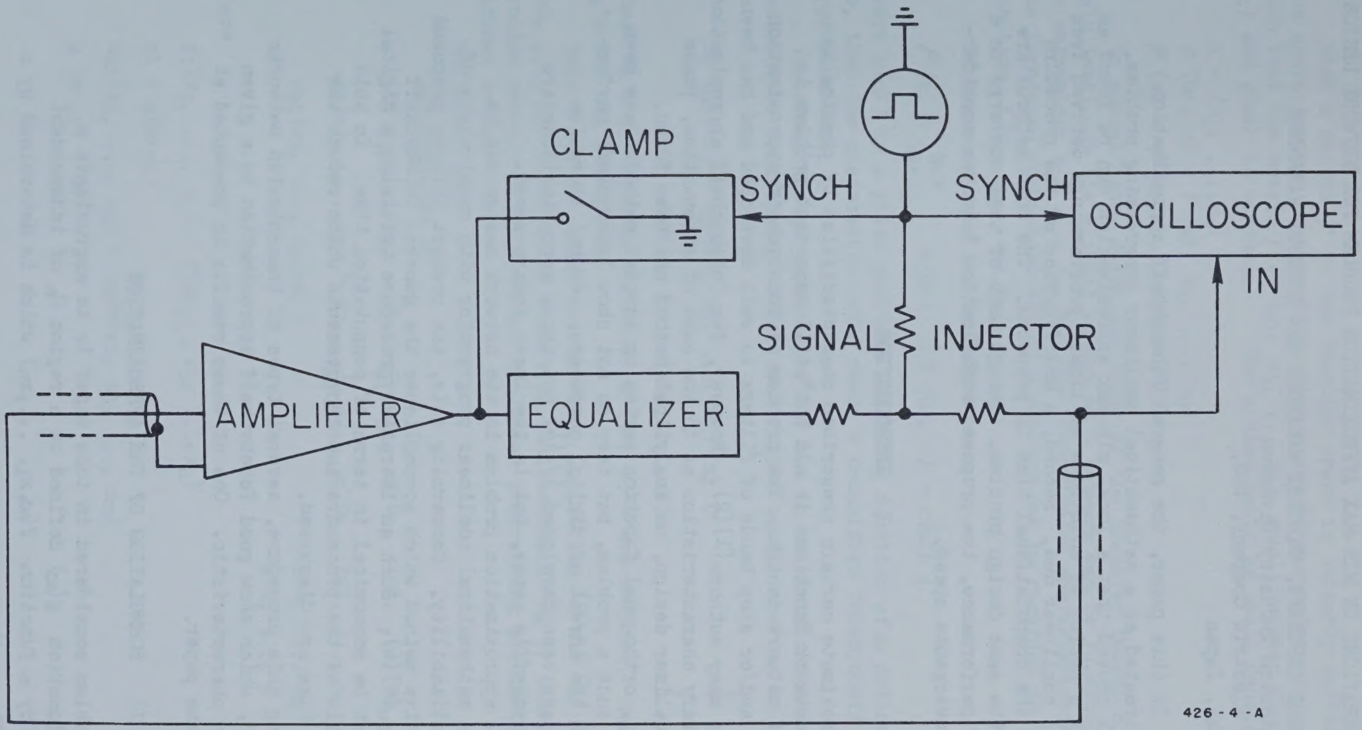

CLAMP
SYNCH
SYNCH
OSCILLOSCOPE
IN
SIGNAL
INJECTOR
AMPLIFIER
EQUALIZER
426 - 4 - A

Figure 3

SOME CONSIDERATIONS ON MIN-MAX APPROXIMATION PROBLEM FOR NETWORK DESIGN

YASUTOSHI ISHIZAKI, HIROSHI WAKAZUKI AND HITOSHI WATANABE
Transmission Industry Division
Nippon Electric Company, Ltd.
Kawasaki, Japan

Abstract - In this paper, the general Tchebycheff approximation problem is treated as a mathematical nonlinear programming problem. First, it is proved that a local min-max approximation can be found as the limit of a series of solutions of linear programmings derived from the original nonlinear one. Second, a modification of the procedure which saves the computational time is proposed. The both methods are applied to the same design problem. Though both of them converge to a Tchebycheff performance, the proposed modification improves considerably the convergence speed.

I INTRODUCTION

To approximate certain prescribed characteristics by physically realizable network functions is old but still important problem in the field of network design. The problem of Tchebycheff approximation in the pass and/or stop bands of filters is well developed and has been described by many authors(1)(2). However, for Tchebycheff approximation to an arbitrary characteristics as in the case of attenuation, phase or delay equalizer design, no analytical method has been found.

Usually, orthogonal function series or templet methods have been applied for such a problem, but they do not show Tchebycheff performance. With the advent of digital computers, several iterative procedures have been developed,(3)(4) but those approximations are not in Tchebycheff's sense, but in the least square sense.

General approximation problem in the network design can be treated as a mathematical nonlinear programming with constraints of physical realizability. Concerning this, the present authors proposed a new iterative method which accomplishes the general Tchebycheff approximation(5)(6). Such an iterative procedure utilizing a digital computer must be economical in terms of computation time. In this paper, details of the proceudre and improvements which reduce the computation time are discussed.

Applying this procedure, several types of transmission networks are designed, which show good Tchebycheff approximation to a given transmission characteristic. One of these results is presented at the end of the paper.

II FORMULATION OF THE APPROXIMATION

The problem considered in this paper is to approximate a prescribed function $g(\omega)$ defined on a region Ω_0 of independent variable ω by a function $f(\omega, x_1, \ldots, x_n)$ which is determined by n parameters $x_1, \ldots, x_n$.

When a metric ρ in the function space is suitably defined, the approximation problem is to find a set of parameters $x_1, \ldots, x_n$ such that the metric ρ (f $(\omega, x_1, \ldots, x_n)$, $g(\omega)$) between $f(\omega, x_1, \ldots, x_n)$ and $g(\omega)$ is minimum. For a simplicity we shall denote

$x = (x_1, \ldots, x_n)$

$f(\omega, x) = f(\omega, x_1, \ldots, x_n)$

$\rho(x) = \rho(x_1, \ldots, x_n) = \rho(f, g)$

In actual case, zeros and poles of the transfer function, the gain level, or the element values of a network can be regarded as parameters $x_1, \ldots, x_n$, and the transmission characteristic or the impedance characteristic of a network can be regarded as function $f(\omega, x_1, \ldots, x_n)$.

If the metric $\rho(x)$ is defined as

$$\rho(x) = \max_{\omega \in \Omega_0} \quad w(\omega) \quad \left| f(\omega, x) - g(\omega) \right| \tag{1}$$

where $w(\omega)$ is a given psoitive weighting function of ω defined on Ω_0, then the minimization of ρ means a nonuniform Tchebycheff approximation.

In actual design problems, $\rho(x)$ has to be minimized under some restrictive conditions. These restrictive conditions can be expressed as

$$\left. \begin{array}{l} h_k(\omega, x) = h_k(\omega, x_1, \ldots, x_n) \leqslant 0 \\ \qquad\qquad \omega \in \Omega_k \qquad (k=1, \ldots, K) \end{array} \right\} \tag{2}$$

where Ω_k's are domains in which $h_k(\omega, x)$'s are defined. Some of h_k's may not contain ω.

Take a set of sample points $\omega_{k1}, \ldots, \omega_{kmk}$ $(\omega_{ki} \in \Omega_k)$ $(k=0, \ldots, K)$ in stead of observing entire domains Ω_k's, then, the problem becomes a mathematical nonlinear programming problem as defined below.

$$\left. \begin{array}{ll} \text{Minimize } x_{n+1} & \\ \quad \text{constrainted by} & \\ \quad f_i(x_1, \ldots, x_n) - g_i - \frac{1}{w_i} x_{n+1} \leqslant 0 & i=1, \ldots, m_0 \\ \quad -f_i(x_1, \ldots, x_n) + g_i - \frac{1}{w_i} x_{n+1} \leqslant 0 & i=1, \ldots, m_0 \\ \quad h_{ki}(x_1, \ldots, x_n) \leqslant 0 & i=1, \ldots, m_k \\ & k=1, \ldots, K \end{array} \right\} \tag{3}$$

where

$f_i(x_1, \ldots, x_n) = f(\omega_{0i}, x_1, \ldots, x_n)$

$g_i = g(\omega_{0i})$

$h_{ki}(x_1, \ldots, x_n) = h_k(\omega_{ki}, x_1, \ldots, x_n)$

$w(\omega_{0i}) = w_i$

III A METHOD FOR SOLVING THE NONLINEAR PROGRAMMING PROBLEM

A number of methods have been proposed for solving nonlinear programming problem. Most of the methods lead to the global minimum (or maximum) only if the set of feasible solutions is convex$^{(7)}$. Otherwise, only a local extreme value can be found. Since, in the problem in the preceding section, the set of feasible solutions is not always convex, we aim at obtaining an extreme value of ρ instead of the global minimum. At the point where $\rho(x)$ has an extreme value, n+1 inequalities among the set of inequalities (3) are fulfilled at the boundary. This means the approximation in Tchebycheff's sense. In the following arguments, an iterative method for finding an extreme value of ρ by means of solving a sequence of linear programming problems associated with the original nonlinear programming problem is presented.

Here, it is assumed that there is no restrictive conditions on x_j's except the ones that x_j's do not change their sign during the process of iteration. In such cases, the nonlinear programming problem corresponding to (3) becomes as follows.

$$\left.\begin{array}{ll}\text{Minimize } x_{n+1} & \\ \quad\text{constrainted by} & \\ \quad f_i(x_1, \ldots, x_n) - g_i - \frac{1}{w_i} x_{n+1} \leqslant 0 & i=1, \ldots, m_0 \\ \quad -f_i(x_1, \ldots, x_n) + g_i - \frac{1}{w_i} x_{n+1} \leqslant 0 & i=1, \ldots, m_0 \\ \quad \frac{x_j}{x_j^0} \geqslant 0 & j=1, \ldots, n \end{array}\right\} \qquad (4)$$

For arbitrary choise of $x_1^0, \ldots, x_n^0$, and $x_{n+1}^0 = \max\limits_{1\leqslant i\leqslant m_0} w_i \left| f_i(x_1^0, \ldots, x_n^0) - g_i \right|$, $x_1^0, \ldots, x_{n+1}^0$ satisfy the constraints of (4). Hence $x_1^0, \ldots, x_{n+1}^0$ can be taken as an initial feasible solution. Beginning at the point $x^0=(x_1^0, \ldots, x_{n+1}^0)$, we proceed to the sequence of points $x^1, x^2, \ldots$. The point x^{r+1} is determined from the point x^r using the solution of a linear programming which is the first order approximation of (4) at x^r. This linear programming problem to be solved is as follows.

$$\left.\begin{array}{l}\text{Minimize } \xi_{n+1} \\ \quad\text{constrainted by} \\ \quad w_i \sum\limits_{j=1}^{n} \frac{1}{x_j} \frac{\partial f_i(x^r)}{\partial x_j} \xi_j - \xi_{n+1} + w_i(f_i(x^r) - g_i) \leqslant 0 \\ \qquad\qquad\qquad\qquad i = 1, \cdots, m_0 \\ \quad -w_i \sum \frac{1}{x_j} \frac{\partial f_i(x^r)}{\partial x_j} \xi_j - \xi_{n+1} - w_i(f_i(x^r) - g_i) \leqslant 0 \\ \quad -\xi_j \leqslant 1 \end{array}\right\} \qquad (5)$$

where $x^r = (x_1^r, \ldots, x_{n+1}^r)$

Linear programming problem (4) can be solved by the well known simplex method, and as a result, ξ^r_{n+1}, the minimum value of ξ_{n+1} is obtained at the point $\overline{\xi}^r = (\overline{\xi}_1^r, \ldots, \xi_n^r)$. When $\overline{\xi}^r \neq 0$, there exists such a positive number α that

$$\rho(x^r(1 + \alpha\overline{\xi^r})) < \rho(x^r) = \rho^r \qquad 0 < \alpha \leqslant 1 \qquad (6)$$

is valid.

To prove this, consider the set of points $(\xi_1, \xi_2, \ldots, \xi_n, \xi_{n+1})$'s, then it is easily realized the subset of points which satisfy the constraints of (4) becomes convex. Clearly, both $\Xi_0 = (0, \ldots, 0, \rho^r)$ and $\Xi_1 = (\overline{\xi}^r, \ldots, \overline{\xi}^r_n, \overline{\xi}^r_{n+1})$ belong to this convex subset. Hence, for $0 < \alpha \leqslant 1$, $(1 - \alpha)\Xi_0 + \alpha\Xi_1 = (\alpha \overline{\xi_1}^r, \ldots, \alpha \overline{\xi_n}^r, \rho^r + \alpha \overline{\Delta\rho^r})$ should satisfy the constraints of (5), where $\overline{\Delta\rho^r} = \overline{\xi}^r_{n+1} - \rho^r$. On the other hand, the constraints of (4) differ from those of (5) by higher order terms than ${\xi^r_j}^2$, there exists such a positive number M that the following inequalities hold.

$$|f_i(x^r(1+\alpha\overline{\xi^r})) - f_i(x^r) - <\alpha\overline{\xi^r}, \nabla f_i(x^r)>| < \alpha^2 M \qquad i = 1, \cdots, m_0 \tag{7}$$

where $< \;\; >$ indicates the inner product. From (5) and (7)

$$f_i(x^r(1+\alpha\overline{\xi^r})) - g_i - \frac{\rho^r - \alpha\overline{\Delta\rho^r}}{w_i} - \alpha^2 M \leqslant 0 \;, \quad -f_i(x^r(1+\alpha\overline{\xi^r})) + g_i - \frac{\rho^r - \alpha\overline{\Delta\rho^r}}{w_i} - \alpha^2 M \leqslant 0 \tag{8}$$
$$i = 1, \cdots, m_0$$

Put $\rho^{r*} = \max_{1 \leqslant i \leqslant m_0} \{ \rho^r + \alpha\overline{\Delta\rho^r} + \alpha^2 M w_i \}$

then for $0 \leqslant \alpha \leqslant 1$, $(x^r(1+\alpha\overline{\xi^r}), \rho^{r*})$ satisfies the constraints of nonlinear programming problem (4). Because $\overline{\xi^r}$ is a solution of (4) $\overline{\Delta\rho^r} < 0$. Therefore, there exists such a positive number α that $\rho^{r*} < \rho^r$. (9)

Now, fix x at $x^h + \alpha\overline{\xi^r}$, then, it can be stated that the minimum value of ξ_{n+1} is not greater than ρ^{r*}. This minimum value becomes

$$\rho^r(\alpha) = \max_{1 \leqslant i \leqslant m_0} \{ w_i | f_i(x^r(1+\alpha\overline{\xi^r})) - g_i | \} \tag{10}$$

Let $\overline{\alpha}$ be the value of α where $\rho^r(\alpha)$ takes its minimum,

$$\rho(\overline{\alpha}) = \min_{0 \leqslant \alpha \leqslant 1} \rho^r(\alpha) \tag{11}$$

and define x^{r+1} and ρ^{r+1} as follows.

$$x^{r+1} = x^r(1+\overline{\alpha}\,\overline{\xi^r}), \qquad \rho^{r+1} = \rho^r(\overline{\alpha}) \tag{12}$$

then from (9), (10), and (12)

$$\rho^{r+1} < \rho^r \tag{13}$$

This proves the statement.

Starting with x^0, and leading to x^1, x^2, ... by repeating this process, the value of ρ can always be decreased so far as $\overline{\xi}^r \neq 0$.

When $\overline{\xi}^r = 0$, the value of ρ can not be decreased any more. In such a case, among the inequalities of (4), n+1 equalities hold at $\xi_j = 0$ for j=1, ..., n. Hence a Tchebycheff approximation is obtained.

IV TRANSFORMATION OF PARAMETERS

If n+1 parameters, $x_1, \ldots, x_n, x_{n+1}$ are expressed as differentiable functions of new parameters, $y_1, \ldots, y_n, y_{n+1}$:

$$\begin{aligned} x_j &= x_j(y_1, \cdots, y_n) \qquad j = 1, \cdots, n \\ x_{n+1} &= y_{n+1} \end{aligned} \tag{14}$$

then, the nonlinear programming problem discussed above, can be transformed into an equivalent problem in n+1-space of y.

Corresponding to linear programming problem (5), the following linear programming problem associated with the nonlinear programming problem of parameters y_j's is obtained.

Minimize η_{n+1} constrainted by

$$w_i \sum_{k=1}^{n} \frac{1}{y_k} \frac{\partial f_i}{\partial y_k} \eta_k - \eta_{n+1} + w_i(f_i(x^r) - g_i) \leqslant 0$$

$$\left.\begin{array}{ll} -w_i\sum_{k=1}^{n}\frac{1}{y_k}\frac{\partial f_i}{\partial y_k}\eta_k-\eta_{n+1}-w_i\left(f_i(x^r)-g_i\right)\leqslant 0 & \\ -\sum_{k=1}^{n}a_k^j\eta_k\leqslant 1 & \begin{array}{l} i=1,\cdots,m_0 \\ j=1,\cdots,n\end{array}\end{array}\right\}\quad(15)$$

where $a_k^j=\frac{y_k}{x_j}\frac{\partial x_j}{\partial y_k}$

Since $\frac{\partial f_i}{\partial y_k}=\sum_{j=1}^{n}\frac{\partial f_i}{\partial x_j}\frac{\partial x_j}{\partial y_k}$ the solution of (5) $(\overline{\xi_1^r},\ldots,\overline{\xi_n^r},\overline{\xi_{n+1}^r})$ is linearly related to the solution of (14), $(\overline{\eta_1^r},\ldots,\overline{\eta_n^r},\overline{\eta_{n+1}^r})$, as shown below:-

$$\begin{bmatrix}\overline{\xi_1^r}\\ \vdots\\ \overline{\xi_n^r}\\ \overline{\xi_{n+1}^r}\end{bmatrix}=\begin{bmatrix}a_1^1 & \cdots & a_n^1 & 0\\ \vdots & & \vdots & \vdots\\ \vdots & & \vdots & \vdots\\ a_1^r & \cdots & a_n^n & 0\\ 0 & \cdots & 0 & 1\end{bmatrix}\begin{bmatrix}\overline{\eta_1^r}\\ \vdots\\ \overline{\eta_n^r}\\ \overline{\eta_{n+1}^r}\end{bmatrix}\qquad(16)$$

The solution of (5) $(\overline{\xi_1^r},\ldots,\overline{\xi_n^r},\overline{\xi_{n+1}^r})$ can immediately be calculated from the solution of (15), and in this sense, (15) is equivalent to (5). But, in general, functions x_j $(y_1,\ldots,y_n)$'s are not linear, so the value α^r for (15) may be different to that of (11). Hence the convergence property of the procedure may become different. For instance, suppose the equations (14) express the relation between the set of element values of two equivalent networks. The convergence speed starting from one equivalent network must be differ from another. Therefore, there is a possibility to reduce the number of iterations of the procedure by utilizing a pertinent equivalent transformation of the network.

In order to reduce the computational time, the transformation of parameters (14) is also effective as discussed in the following section.

V MODIFICATION OF THE PROCEDURE

Many actual design examples indicate us the average ratio of computation time of each step of the procedure described in section III as shown in table I.

I	computation of derivatives $\frac{\partial f_i}{\partial x_j}$	56 %
II	solution of linear programming problem (5)	16 %
III	computation of $\overline{\alpha^r}$	28 %

Table I Average ratio of computation time

This shows that the computation time for the partial derivatives $\frac{\partial f_i}{\partial x_j}$'s is dominant, so in order to save the total computation time, it is necessary to reduce the time for part I.

For this purpose, the following transformation of parameters and an approximation for calculation of partial derivatives $\frac{\partial f_i}{\partial y_k}$'s are applied. Let ξ^h be the solution of local linear programming obtained from original problem at point x^h. If $\overline{\xi^{r_1}},\ldots,\overline{\xi^{r_n}}$ $(r_1\leqslant r,\ \ldots,\ r_n\leqslant r)$ are linearly independent, x_i's can be expressed as linear combinations of these vectors:-

$$x_i=\sum_{k=1}^{n}y_k\,\overline{\xi^{r_k}}\qquad(17)$$

or in a matrix form.

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \overline{\xi_1^{r_1}} & \cdots & \overline{\xi_1^{r_n}} \\ \vdots & \ddots & \vdots \\ \overline{\xi_n^{r_1}} & \cdots & \overline{\xi_n^{r_n}} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

In this case, $x_j(y_1, \dots, y_n)$'s in(14) are linear functions of y_k's.

In the process from x^h to $x^{h+1}=x^h(1+\overline{\alpha^h}\,\overline{\xi^h})$ for $h<r$, $\overline{\alpha^h}$ is so determined that the maximum of the values of $w_i\left|f_i(x^h(1+\alpha\overline{\xi^h}))-g_i\right|$'s $(i=1, \dots, m_0)$ may be minimum. Therefore, $w_i(f_i - g_i)$'s must be computed for several values of α. Taking account of these values, the values of $\frac{\partial f_i}{\partial y_k}$'s can be approximated by

$$w_i\,\frac{\partial f_i(x^{r_k})}{\partial y_k} = \frac{w_i\left[\left(f_i(x^{r_k}(1+\alpha_1\overline{\xi^{r_k}}))-g_i\right)-\left(f_i(x^{r_k}(1+\alpha_2\overline{\xi^{r_k}}))-g_i\right)\right]}{\alpha_1^{r_k}-\alpha_2^{r_k}} \quad i=1,\dots,m_0 \qquad (19)$$

where $\alpha_1^{r_k}$ and $\alpha_2^{r_k}$ are the two values of α next to $\overline{\alpha^{r_k}}$ which are used in the process from x^{r_k} to x^{r_k+1}. The partial derivatives thus obtained are not the exact values of them at the point x^r, but at the point $x^{r_1}, \dots, x^{r_k}$. Hence, it is not always assured that $0<\overline{\alpha^r}\leqslant 1$. In some cases, $\overline{\alpha^r}$ may be zero.

Since, in such cases, $\frac{\partial f_i(x^{r_k})}{\partial y_k}$'s are not sufficient approximations for $\frac{\partial f_i(x^r)}{\partial y_k}$, we must compute $\frac{\partial f_i}{\partial x_j}$'s at the point x^r, and start again whole process described above.

VI ACTUAL COMPUTATIONAL METHOD

Whole computer program is summarized in the following algorithm and the corresponding flow chart is given in Fig. 1.

On the first round of the iterative procedure, all partial derivatives, $\frac{\partial f_i}{\partial x_j}$'s must be computed and the next point x^1 is obtained by the procedure described in section III. If we put $y_j=x_j\ (j=1, \dots, n)$ and $(a_k^j) = I_n$, where I_n is the unit matrix of order n, then the procedure of section V will become identical with that of section III. Hence no special computer program is needed for the first one round of the iteration.

[1] Compute $w_i\,\frac{f_i(x_1, \dots, x_j+\Delta x_j, \dots, x_n) - f_i(x_1, \dots, x_n)}{\Delta x_j}$'s $(i=1,\dots, m_0,\ j=1, \dots, n)$ as approximations for the partial derivatives $w_i\frac{\partial f_i}{\partial x_j}$'s, and $w_i(f_i(x_1,\dots,x_n)-g_i)$'s $(i=1,\dots, m_0)$. Store them in the form of the simplex tableau of linear programming (5).

[2] Make (a_k^j) unit matrix.

[3] Solve linear programming problem (15) and find the solution $\overline{\eta^r}$. For (15) a modified simplex method must be applied because there is no nonnegative condition for η_k's in (15).

[4] Calculate $\overline{\xi^r} = (a_k^j)\,\overline{\eta^r}$.

[5] If $\overline{\xi^r}=0$, Tchebycheff approximation is obtained. Hence the iterative procedure is finished. Otherwise go to the next step.

[6] Find $\overline{\alpha^r}$ that minimizes $\rho^r(\alpha)=\max\limits_{1\leqslant i\leqslant m_0}\left[w_i\left|f_i(x^r(1+\alpha\overline{\xi^r}))-g_i\right|\right]$ for $0<\alpha\leqslant 1$, and define x^{r+1} by (12).

In our program, how to find $\overline{\alpha^r}$ is as follows.

Calculate the values of $\max\limits_{1\leqslant i\leqslant m_0}\left[w_i\left|f_i(x^r(1+\alpha\overline{\xi^r})-g_i\right|\right]$'s for $\alpha=0, \frac{1}{8}, \frac{2}{8}, \dots, 1$, and find α_1 such that

$$\max_{1\leqslant i\leqslant m_0}\left[w_i\left|f_i(x^r(1+\alpha_1\overline{\xi^r}))-g_i\right|\right]\leqslant\max_{1\leqslant i\leqslant m_0}\left[w_i\left|f_i(x^r(1+(\alpha_1+\tfrac{1}{8})\overline{\xi^r}))-g_i\right|\right]$$

If there is no such an $\alpha_1 (0 \leqq \alpha_1 < 1)$, put $\alpha_1=1$. Then there are three possible cases $\alpha_1=0$, $0<\alpha_1<1$ and $\alpha_1=1$.

a) If $\alpha_1=0$, calculate

$$\max_{1 \leqq i \leqq m_0} \left[w_i \left| f_i(x^r(1+\alpha\overline{\xi^r})) - g_i \right| \right] \quad \text{for } \alpha = \frac{1}{64}, \frac{2}{64}, \cdots, \frac{1}{8}$$

and find such an α_2 that

$$\max_{1 \leqq i \leqq m_0} \left[w_i \left| f_i(x^r(1+\alpha_2\overline{\xi^r})) - g_i \right| \right] < \max_{1 \leqq i \leqq m_0} \left[w_i \left| f_i(x^r(1+(\alpha_2+\tfrac{1}{64})\overline{\xi^r})) - g_i \right| \right]$$

then put $\overline{\alpha^r}=\alpha_2$

b) If $\alpha_1 \neq 0$, $\alpha_1 \neq 1$, calculate

$$\max_{1 \leqq i \leqq m_0} \left[w_i \left| f_i(x^r(1+\alpha\overline{\xi^r})) - g_i \right| \right] \quad \text{for } \alpha = \alpha_1 - 1,\ (\alpha_1-1)+\frac{1}{32},$$

$(\alpha_1-1)+\frac{2}{32}, \ldots$ and find such an α_2 that

$$\max_{1 \leqq i \leqq m_0} \left[w_i \left| f_i(x^r(1+\alpha_2\overline{\xi^r})) - g_i \right| \right] < \max_{1 \leqq i \leqq m_0} \left[w_i \left| f_i(x^r(1+(\alpha_2+\tfrac{1}{32})\overline{\xi^r})) - g_i \right| \right)$$

then, put $\overline{\alpha^r}=\alpha_2$

c) If $\alpha_1=1$, divide the interval $(\frac{7}{8}, 1)$ of α and find $\overline{\alpha^r}$ by the same process as in the case of $\alpha_1=0$.

This process is illustrated in Fig. 2.

[7] If $\overline{\alpha^r}=0$, the partial derivatives $w_i \dfrac{\partial f_i(x^r)}{\partial x_j}$'s must be computed. Hence go to [1], otherwise go to [8].

[8] Let $x^r(1+\overline{\alpha^r \xi^r})$ be new values of the parameters x^{r+1}.

[9] Replace one of the columns of the matrix $w_i(\dfrac{\partial f_i}{\partial y_{rk}})$ ($i=1, \ldots, m_0$, $k=1, \ldots, n$) by (19). The column to be replaced, say qth column, is determined by the condition

$$\frac{\left[\sum_{i=1}^{m_0} w_i \dfrac{\partial f_i}{\partial y_r} \cdot w_i \dfrac{\partial f_i}{\partial y_{r_k}} \right]^2}{\sum_{i=1}^{m_0} \left[w_i \dfrac{\partial f_i}{\partial y_{r_k}} \right]^2} = \min. \quad \text{for } r_k = q,$$

where $w_i \dfrac{\partial f_i}{\partial y_r}$ is approximated by (19).

[10] Replace the qth column of the matrix (a_k^i) by $(\xi_1^r, \ldots, \xi_n^r)'$, where prime indicates transpose of a matrix.

The processes [3] to [9] are repeated until Tchebycheff approximation is obtained or the number of iterations r exceeds some predetermined number N.

VII EXAMPLE

The method described in this paper is applied to various sorts of transmission network designs and turns out to be effective for Thebycheff approximation of any kind of networks such as fixed and variable equalizers, phase equalizers, and filters. One design example shall be illustrated in this section.

As an experiment, two calculation methods described in section III and V are applied to the same design problem.

Fig. 3 shows the maximum deviation, ρ^r, versus number of iterations r for both methods. It can be observed that at the initial several rounds of iteration where the approximation is poor, the procedure of section V is not effective. So it is better to compute the partial derivatives at each round in the initial stage of iteration.

This comes from the fact that the values of parameters $(x_1, \ldots, x_n)$ vary considerably on each round in initial stage, and hence derivatives along the direction of vector $\overline{\xi^r}_k$ at point x^{r_k} are not sufficient approximations for those values at the correct point x^r.

But after the deviation ρ^r becomes small, the procedure of section V is very effective.

To illustrate the computation times of two procedures, maximum deviation versus computation time is shown in Fig. 4, which shows the proposed method saves the computation time considerably.

VIII ACKNOWLEDGMENT

The authors wish to express their thanks to Prof. Masao Iri of the University of Tokyo for his constructive discussions and excellent suggestions throughout this work. They are also grateful to Mr. Masanori Kogo, Manager of Components Engineering Department, and the members of Network Section of Nippon Electric Co. for their support.

REFERENCES

(1) S. Darlington, "The Potential Analogue Method of Network Synthesis" BSTJ April 1951 pp. 315 ~ 365.

(2) H. Watanabe, "Approximation Theory for Filter Network" IRE Trans. on Circuit Theory Vol. CT-9, No.3, pp. 341 ~ 356 Sept. 1961.

(3) M. R. Aaron, "The Use of Least Squares in System Design" IRE Trans. Vol. CT-3 No. 4, December 1956.

(4) C. A. Desoer, S. K. Mitra, "Design of Lossy Ladder Filters by Digital Computer" IRE Trans. Vol. CT-9, No.3, Sept. 1961.

(5) Y. Ishizaki, H. Watanabe, "On an Iterative Method for Network Design — A Consideration as a Nonlinear Programming Problem —." (in Japanese) P.T.G. on Circuit Theory Japan. May 1965.

(6) Y. Ishizaki, H. Watanabe, "An Iterative Method for Network Design as a Nonlinear Programming Problem" Monograph DEB 485, Internal Report, Nippon Electric Company Ltd. Presented at the I.C.M.C.I. Sept. 7-11 1964, Tokyo Japan.

(7) P. Wolfe, "Method of Nonlinear Programming" Recent Advances in Mathematical Programming. McGraw - Hill.

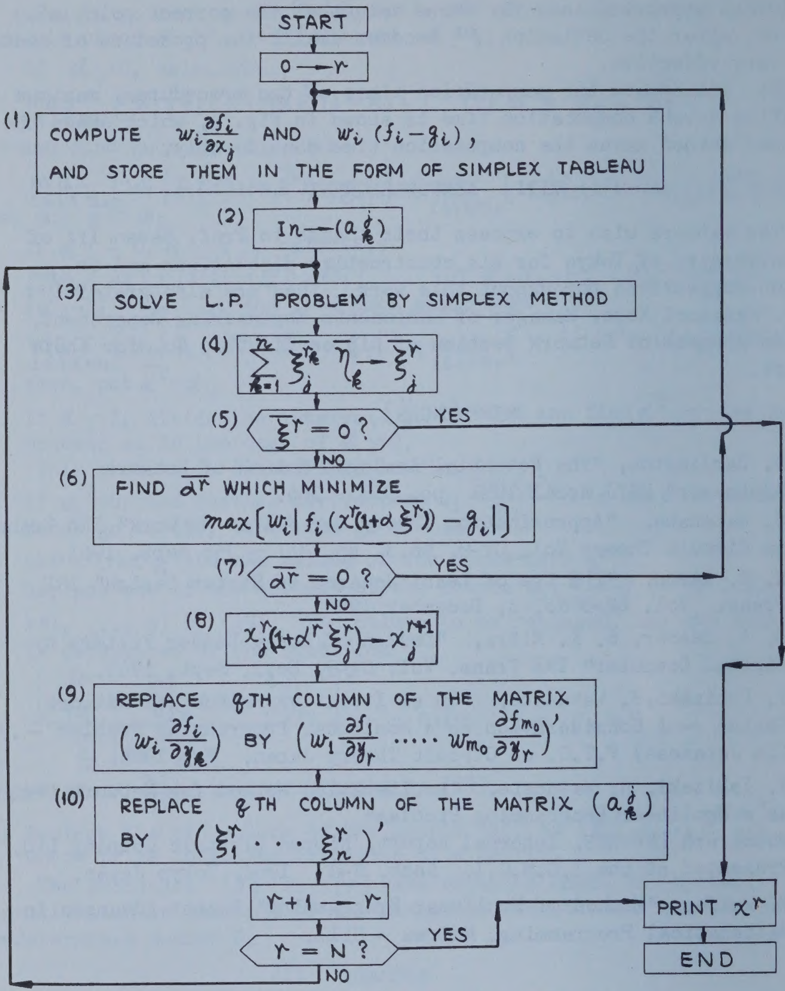

START
$0 \to r$
(1) COMPUTE $w_i \frac{\partial f_i}{\partial x_j}$ AND $w_i(f_i - g_i)$, AND STORE THEM IN THE FORM OF SIMPLEX TABLEAU
(2) $In \to (a^j_k)$
(3) SOLVE L.P. PROBLEM BY SIMPLEX METHOD
(4) $\sum_{k=1}^{n} \overline{\xi}^{r_k}_j \overline{\eta}_k \to \overline{\xi}^r_j$
(5) $\overline{\xi}^r = 0$?
YES
NO
(6) FIND $\overline{\alpha}^r$ WHICH MINIMIZE $\max[w_i | f_i(x^r(1+\alpha\overline{\xi}^r)) - g_i|]$
(7) $\overline{\alpha}^r = 0$?
YES
NO
(8) $x^r_j(1+\overline{\alpha}^r\overline{\xi}^r_j) \to x^{r+1}_j$
(9) REPLACE qTH COLUMN OF THE MATRIX $\left(w_i \frac{\partial f_i}{\partial y_k}\right)$ BY $\left(w_1 \frac{\partial f_1}{\partial y_r} \cdots\cdots w_{m_0} \frac{\partial f_{m_0}}{\partial y_r}\right)'$
(10) REPLACE qTH COLUMN OF THE MATRIX (a^j_k) BY $(\xi^r_1 \cdots \xi^r_n)'$
$r+1 \to r$
$r = N$?
YES
NO
PRINT x^r
END

FIG. 1 FLOW CHART OF THE PROCEDURE

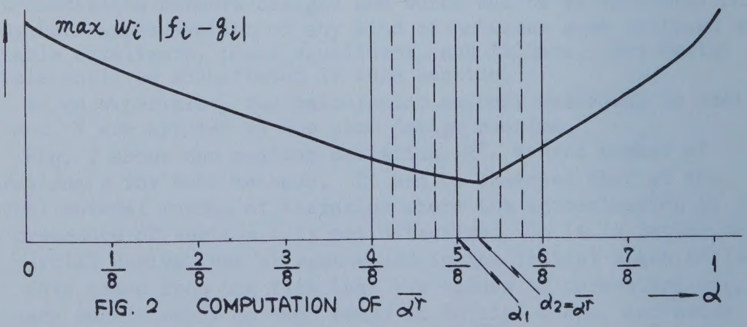


FIG. 2 COMPUTATION OF $\overline{\alpha}^r$

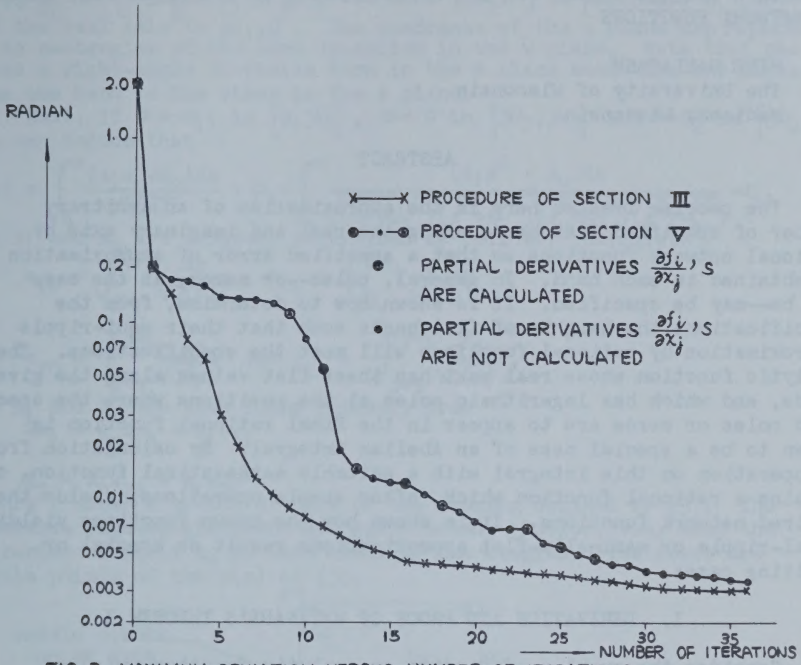


FIG. 3 MAXIMUM DEVIATION VERSUS NUMBER OF ITERATIONS

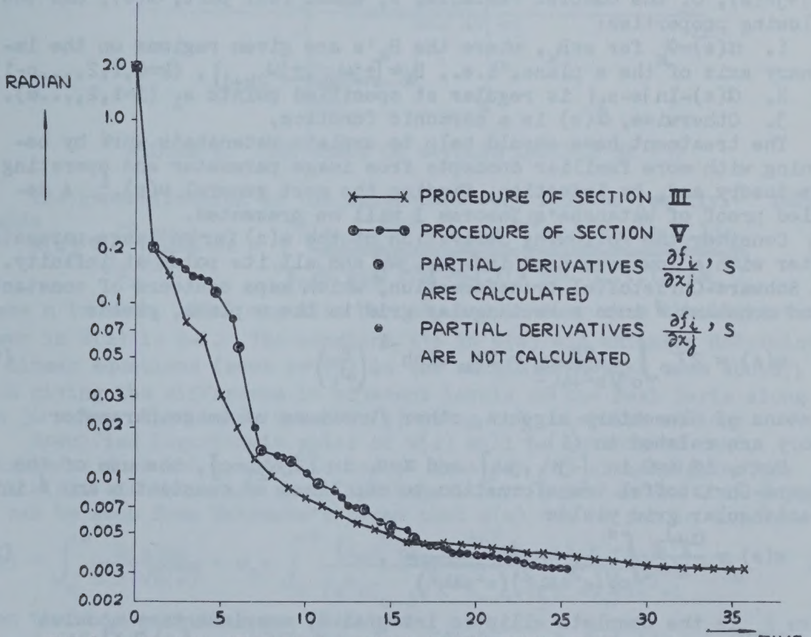


FIG. 4 MAXIMUM DEVIATION VERSUS COMPUTATION TIME

TOWARDS A UNIFIED THEORY FOR THE APPROXIMATION OF SPECIFIED FLAT BANDS BY NETWORK FUNCTIONS

MIKE SABLATASH
The University of Wisconsin
Madison, Wisconsin

ABSTRACT

The problem treated here is the approximation of an arbitrary number of specified flat bands along the real and imaginary axis by rational network functions so that a specified error of approximation is obtained in each band. In general, poles--or zeros, as the case may be--may be specified. It is shown how to determine, from the specifications, the heights of flat bands such that their equi-ripple approximation by rational functions will meet the specifications. The analytic function whose real part has these flat values along the given bands, and which has logarithmic poles at the positions where the specified poles or zeros are to appear in the final rational function is shown to be a special case of an Abelian integral. By calculation from, or operation on this integral with a suitable mathematical function, one obtains a rational function which, after simple operations, yields the desired network functions. It is shown how the known functions yielding equal-ripple or maximally-flat approximations result as special or limiting cases.

I. DERIVATION AND PROOF OF WATANABE'S THEOREM I

Consider the problem of finding an analytic function, $w(s) = \alpha(s)+j\beta(s)$, of the complex variable, s, whose real part, $\alpha(s)$, has the following properties:

1. $\alpha(s)=\alpha_k$ for $s\epsilon B_k$, where the B_k's are given regions on the imaginary axis of the s plane, i.e., $B_k=[\pm j\omega_{2k},\pm j\omega_{2k+1}]$, $(k=0,1,2,\ldots n-1)$.
2. $\alpha(s)-\ln|s-s_i|$ is regular at specified points s_i $(i=1,2,\ldots m)$.
3. Otherwise, $\alpha(s)$ is a harmonic function.

The treatment here should help to explain Watanabe's work by beginning with more familiar concepts from image parameter and operating loss theory and, by induction, finding the most general w(s).[1] A detailed proof of Watanabe's Theorem I will be presented.

Consider the following derivation of the w(s) for an image-parameter filter with a low-pass band in $[-j\omega_1, j\omega_1]$ and all its poles at infinity. The Schwarz-Christoffel transformation, which maps contours of constant α and constant β into a rectangular grid in the w plane, yields

$$w(s) = 2\gamma_\infty \int_0^s \frac{ds}{\sqrt{s^2+\omega_1^2}} = 2\gamma_\infty \sinh^{-1}\left(\frac{s}{\omega_1}\right). \tag{1}$$

By means of elementary algebra, other functions of image-parameter theory are related to (1).[2]

Next, if $\alpha=0$ in $[-j\omega_1, j\omega_1]$ and $\alpha=\alpha_1$ in $[\pm j\omega_2,\infty]$, the use of the Schwarz-Christoffel transformation to map lines of constant α and β into a rectangular grid yields

$$w(s) = \frac{\alpha_1\omega_2}{K'}\int_0^s \frac{ds}{\sqrt{(s^2+\omega_1^2)(s^2+\omega_2^2)}}, \tag{2}$$

where K' is the complete elliptic integral of complementary modulus $k'=\sqrt{1-k^2}$, where $k=\omega_1/\omega_2$. The band $[0, j\omega_1]$ maps to $\left[0, j\frac{\alpha_1 K}{K'}\right]$, the band

$[j\omega_1, j\omega_2]$ to $\left[0+j\frac{\alpha_1 K}{K'}, \alpha_1+j\frac{\alpha_1 K}{K'}\right]$, the band $[j\omega_2, \infty]$ to $\left[\alpha_1+j\frac{\alpha_1 K}{K'}, \alpha_1+j0\right]$ and the real axis to $[\alpha_1, 0]$. The quadrants of the s plane map repeatedly to rectangles of the same dimension in the w plane. Note that one takes a right-angle clockwise turn in the w plane each time one passes from one band to the other in the s plane.

Next, if $\alpha = \alpha_1$, in $[0, j\omega_1]$, $\alpha = 0$ in $[j\omega_2, j\omega_3]$ and $\alpha = \alpha_2$ in $[j\omega_4, \infty]$, one can deduce that

$$w(s) = \int_0^s \frac{(A_1 s^2 + A_0)ds}{\sqrt{Q(s)}} + \alpha_1 = \int_0^s \frac{(A_1 s^2 + A_0)ds}{\sqrt{(s^2+\omega_1^2)(s^2+\omega_2^2)(s^2+\omega_3^2)(s^2+\omega_4^2)}} + \alpha_1. \quad (3)$$

The A_1 and A_0 are uniquely determined as follows. Calculate

$$a_{10} = \int_{j\omega_1}^{j\omega_2} \frac{dp}{\sqrt{Q(p)}}, \quad a_{11} = \int_{j\omega_1}^{j\omega_2} \frac{p^2 dp}{\sqrt{Q(p)}}$$

$$a_{20} = \int_{j\omega_3}^{j\omega_4} \frac{dp}{\sqrt{Q(p)}} \text{ and } a_{21} = \int_{j\omega_3}^{j\omega_4} \frac{p^2 dp}{\sqrt{Q(p)}}.$$

The A_1 and A_0 are then found uniquely from

$$\left.\begin{aligned} -\alpha_1 &= A_0 a_{10} + A_1 a_{11} \\ \alpha_2 &= A_0 a_{20} + A_1 a_{21} \end{aligned}\right\} \quad (4)$$

Figure 1 shows a sketch of part of the mapping defined by (3). The clockwise right-angle turns are again noted. Notice, too, that the character of $A_1 s^2 + A_0$ is made clear by the diagram. Its zeros are saddle points of the w(s) of (3).

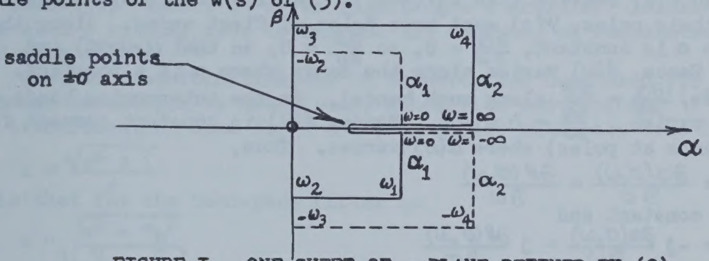


FIGURE I. ONE SHEET OF w-PLANE DEFINED BY (3)

The generalization to the multi-band case follows readily.[3] This yields

$$w(s) = \int_{s_0}^s \frac{N(s)}{\sqrt{D(s)}} ds = \int_{s_0}^s \frac{(A_{n-2}s^{n-2} + A_{n-4}s^{n-4} + \ldots. A_2 s^2 + A_0)ds}{\sqrt{(s^2+\omega_1^2)(s^2+\omega_2^2)(s^2+\omega_3^2)\ldots.(s^2+\omega_n^2)}} \quad (5)$$

where n is even. N(s) may also be odd, and in that case the highest power in N(s) is n-1. The constant A's in N(s) are uniquely determined by linear equations (such as (4) in the simple band-pass case above), each giving the difference in adjacent levels of the real parts along the $j\omega$ axis. Klinkhamer deduced (5) using the potential analogy.

Specified logarithmic poles of w(s) will be introduced next. Klinkhamer does not treat this case. Equation (1) already gives one such w(s). If there are specified logarithmic poles and $\alpha = 0$ in $[-j, j]$, it can be seen from Watanabe's paper that w(s) can be put in the form

$$w(s) = \int_0^s \frac{N(s)ds}{R(s)\sqrt{D(s)}} + w_0 = \int_0^s \frac{(A_m s^m + A_{m-2}s^{m-2} + \ldots + A_2 s^2 + A_0)}{(s^m + C_{m-2}s^{m-2} + \ldots + C_2 s^2 + C_0)\sqrt{s^2+1}} + w_0. \quad (6)$$

When there is more than one flat band

$$w(s) = \int_0^s \frac{(A_{n+m-2}\, s^{n+m-2} + A_{n+m-4}\, s^{n+m-4} + \ldots + A_2 s^2 + A_0)ds}{(s^m + C_{m-2}p^{m-2} + \ldots + C_2 p^2 + C_0)\sqrt{(p^2+\omega_1^2)(p^2+\omega_2^2)\ldots(p^2+\omega_n^2)}} + w_0 \quad (7)$$

where n is even. The C's are determined by the given pole positions, while the A's are found from the (specified) residues at the poles of $w'(s)$ and the differences in levels in the flat bands.

Now, if $R_1(w,z)$ is any rational function of w and z then

$$\int_c R_1(w,z)dz \tag{8}$$

is an Abelian integral for the algebraic function w defined by $P(z,w)=0$ where P is a polynomial in z and w.[4] When this latter relation is $w^2-P_n(z)=0$, where P_n is a polynomial of the nth degree in z, the Abelian integral is named a hyperelliptic integral, for $n>6$. For $n=5$ or 6 it is called ultraelliptic and for $n=3$ or 4 it is an elliptic integral. If the integrand has no poles the integral is, as (5), of the first kind, whereas if the integrand has poles it is, as (7), of the third kind. A detailed proof of Watanabe's Theorem I will now be presented. The theorem is:

The ideal transmission function, w(s), is a hyperelliptic integral of the third kind, and the nonpole ideal transmission function is a hyperelliptic integral of the first kind. They can be expressed as

$$w(s) = \int_{s_0}^{s}\psi(s)ds,$$

where $\psi(s)$ is an algebraic function of s.

Proof: Let

$$\psi(s) = \frac{dw(s)}{ds} = \frac{\partial\alpha(\sigma,\omega)}{\partial\omega} - j\frac{\partial\alpha(\sigma,\omega)}{\partial\omega}. \tag{9}$$

At the m points, s_j, $\psi(s)$ has poles of first order since α behaves like $k_j\ln|s-s_j|$ near $s=s_j$. Therefore, w(s) behaves like $k_j\ln(s-s_j)$ near $s=s_j$, and so $\psi(s)$ behaves like $k_j/(s-s_j)$. Qualitatively, since w(s) has logarithmic poles, $\psi(s)$ must have poles of first order. Along the bands where α is constant, $\frac{\partial\alpha}{\partial\omega} = 0$, so $\frac{\partial\beta}{\partial\sigma} = 0$, so that $\alpha=\alpha(\sigma)$ and $\beta=\beta(\omega)$. Hence, $\beta(\omega)$ varies along the bands where α is a constant. (Furthermore, $\frac{\partial\alpha}{\partial\sigma} = \frac{\partial\beta}{\partial\omega}$ along such bands). In the intervening bands, where $\alpha(\omega)$ varies, $\frac{\partial\alpha}{\partial\sigma} = 0 = \frac{\partial\beta}{\partial\omega}$. Hence, $\beta(\omega)$ is constant (except for discrete jumps at poles) where $\alpha(\omega)$ varies. Thus,

$$\psi(s) = \frac{\partial\alpha(\sigma,\omega)}{\partial\sigma} = \frac{\partial\beta(\sigma,\omega)}{\partial\omega}$$

where α is constant and

$$\psi(s) = -j\frac{\partial\alpha(\sigma,\omega)}{\partial\omega} = j\frac{\partial\beta(\sigma,\omega)}{\partial\sigma}$$

where α varies (β is constant). Thus, $\psi(s)$ changes from purely real to purely imaginary or vice versa at the ends of the flat bands, as one proceeds along the imaginary axis. Therefore, $\psi(s)$ is not analytic at these ends. (Clearly, it is a simple matter to have flat bands along the $\pm\sigma$ axis). The nature of the points $\pm j\omega_k$ ($k=1,2,\ldots.n$) will be established next, so that the analytic behaviour of $\psi(s)$ at all points in the s plane will be specified.

Let the neighborhood of, say, $j\omega_k$, be mapped to a neighborhood of $\psi=\infty$. Let

$$t' = s - j\omega_k \text{ and } t = \psi(s).$$

The angle of t' increases by π when s passes through $j\omega_k$ and $\psi(s)$ changes from purely real to purely imaginary, or vice versa. From the simple examples that were considered the angle of t increases by $\pi/2$, so that $\arg(t') = -2\arg(t)$ in the neighborhood. With the scale factor ϵ one obtains, from the basic theory of conformal mapping,

$$t^2 = \epsilon(t')^{-1}.$$

Hence, the $j\omega_k$ are pole branch points of order two. (This also follows, by induction, from the simple cases considered earlier).

It was shown earlier how the numerator of the $\psi(s)$ is uniquely determined from the specification of the residues at the prescribed poles and the differences in the levels of the flat attenuation bands.

(Changes in β may also be used instead of changes in α). Thus, the required $\psi(s)$'s have now been completely specified as functions of s. Now there is the theorem:[5]

"Every domain of values which is given in any manner 'in the small' generates, if at all, precisely one well-determined monogenic analytic function." In our case the Riemann surface corresponding to the monogenic analytic function is completely determined. Furthermore, the classification of the points on this Riemann surface shows that $\psi(s)ds$ is an Abelian differential of the third kind, and so $w(s)$ is an Abelian--specifically, a hyperelliptic--integral of the third kind.[6,7,8] $\alpha(j\omega)$ is constant where $\psi(j\omega)$ is purely real and $\beta(j\omega)$ is constant where $\psi(j\omega)$ is purely imaginary. The nonpole ideal transmission function is clearly an Abelian integral of the first kind. This completes the proof of the theorem.

One should note that all image propagation functions are special cases of $w(s)$, obtained when the residues are not negative.

II. SOME CONFORMAL TRANSFORMATIONS

Watanabe has used special cases of the following transformation:[9]

$$\frac{c_1-c_o}{c_1-c_2} = \frac{s^2-c_2}{s^2-c_o} = \frac{d_1-d_o}{d_1-d_2}\ \frac{\lambda^2-d_2}{\lambda^2-d_o}\,. \tag{10}$$

The use of special cases of this transformation for low-pass, band-pass and band-stop filters is given in Watanabe's article. Iedokoro, et. al., have used a special case to improve the accuracy of computation of band-pass filters.[10]

Another transformation, due to Bennett, has been used for computational advantages by Szentirmai, Schoeffler and Bingham.[11,12,13,14] The transformation used for a low-pass filter with pass band in $[-j,j]$ is

$$z = \frac{\sqrt{s^2+1}}{s} \tag{11}$$

while that for the band-pass filter is

$$z = \sqrt{\frac{s^2+\eta_p^2}{p^2+\eta_{-p}^2}}, \tag{12}$$

where $\eta_{-p} = \frac{f_{-p}}{f_o}$, $\eta_p = \frac{f_p}{f_o}$ and $f_o = \sqrt{f_{-p}\,f_p}$.

In either case the ideal transmission function in the z plane is given by

$$w(z) = \sum_{i=1}^{N/2} \ln\left[\frac{z+\frac{\gamma_i}{|\gamma_i|}z_i}{z-\frac{\gamma_i}{|\gamma_i|}z_i}\right] \tag{13}$$

$$= \sum_{i=1}^{N/2} \ln\left[\frac{z^2+\frac{2\gamma_i}{|\gamma_i|}z_iz+z_i^2}{z^2-z_i^2}\right], \tag{14}$$

where

$$N = 2|\gamma_\infty| + \sum_{\nu=1}^{h} 2|\gamma_\nu| + \sum_{\mu=1}^{l} 2|\gamma_\mu| + \sum_{i=1}^{g} 4|\gamma_i|, \tag{15}$$

the γ's are the residues (as before) and the z_i are the poles mapped into the z plane according to (11) in the low-pass case and according to (12) in the band-pass case. From (14) it is easy to see that Saal and Ulbrich's use of Bennett's procedure to obtain the numerator of a characteristic function applies when some of the γ's are negative.[15] Thus, the characteristic function is $\phi(z) = \epsilon \cosh w(z)$, and its numerator is clearly an even polynomial in z.

The transformations

$$\Gamma = -\tfrac{1}{2}\ln(1 + \lambda^{-2}) \tag{16}$$

and

$$\Gamma = \tfrac{1}{2}\ln \frac{s^2 + \omega_p^{\,2}}{s^2 + \omega_{-p}^{\,3}} \, \frac{\omega_{-p}}{\omega_p} \tag{17}$$

for low- and band-pass cases, respectively, originated with Rumpelt, who used them in a template method of finding optimum pole positions for image-parameter filters and, later, for reference filters for synthesis by means of the operating-loss method.[16,17] The w(s) becomes

$$w(\Gamma) = \sum_{i=1}^{N/2} \gamma_i \ln \coth \frac{(\Gamma - \Gamma_\nu)}{2} \tag{18}$$

when (16) or (17) are used. This makes it clear that the templates are turned upside down when γ is negative.

III. DERIVATION OF SPECIFICATIONS ON w(s) FROM SPECIFICATIONS OF RIPPLE AND FLAT LOSS FOR FINAL NETWORK

Let $\frac{E(s)}{P(s)}$ be the operating loss and $\phi(s) = \frac{F(s)}{P(s)}$ be the characteristic function for a doubly-terminated lossless filter. Then

$$E(s)E(-s) = F(s)F(-s) + P(s)P(-s). \tag{19}$$

The transducer loss in nepers is

$$A_n = \tfrac{1}{2}\ln[1 + \phi(s)\phi(-s)]. \tag{20}$$

If the flat attenuation has minima of zero and maxima of $A_{n\nu}$ then it is easy to see that $\ln\left|\frac{F}{P}\right|$ varies between $-\infty$ and $\frac{1}{2}\ln(e^{2A_{n\nu}} - 1)$, or, approximately, $\frac{1}{2}\ln 2 + \frac{1}{2}\ln A_{n\nu}$, in the pass bands. If the pass-band minima are equal to $a_{n\nu}$, then if one lets

$$\left|\frac{\overline{F}}{P}\right|^2 = \left|\frac{F}{P}\right|^2 - (e^{2a_{n\nu}} - 1)$$

$\overline{F}$ has its zeros along the $j\omega$ axis and the ripple there is

$$\alpha_\nu = \ln\left|\frac{\overline{F}}{P}\right| = a_{n\nu} + \tfrac{1}{2}\ln\left[e^{2A_{n\nu}} - 1\right].$$

The zeros of F will lie on the contours where $\alpha(s) = \ln\sqrt{e^{2a_{n\nu}} - 1}$, on the same lines of constant phase, β, as the zeros of $\overline{F}$ would be placed in the w plane which will be derived shortly.

The flat stop bands will be considered next. Let the peak-to-peak ripple be $\overline{A}_{n\nu}$. It is easy to see that

$$\left|\frac{P}{F}\right|^2_{\text{min.}} = \frac{1}{(e^{2\bar{a}_{n\nu} + 2\overline{A}_{n\nu}} - 1)} \quad \text{and} \quad \left|\frac{P}{F}\right|^2_{\text{max.}} = \frac{1}{(e^{2\bar{a}_{n\nu}} - 1)}$$

and the ripple of $\left|\frac{P}{F}\right|^2$ is $\dfrac{e^{2\bar{a}_{n\nu}}(e^{2\overline{A}_{n\nu}} - 1)}{(e^{2\bar{a}_{n\nu}} - 1)(e^{2\bar{a}_{n\nu} + 2\overline{A}_{n\nu}} - 1)}$.

The function

$$\left|\frac{\overline{P}}{\overline{F}}\right|^2 = \left|\frac{P}{F}\right|^2 = \frac{1}{(e^{2\bar{a}_{n\nu} + 2\overline{A}_{n\nu}} - 1)}$$

has its zeros along the $j\omega$ axis. The ripple is the same as that of $\left|\frac{P}{F}\right|^2$. Then the minima of $\left|\frac{F}{\overline{P}}\right|^2$, which possesses imaginary-axis poles, is given by

$$\left|\frac{F}{\overline{P}}\right|^2_{\text{min.}} = \frac{(e^{2\bar{a}_{n\nu}} - 1)(e^{2\bar{a}_{n\nu} + 2\overline{A}_{n\nu}} - 1)}{e^{2\bar{a}_{n\nu}}(e^{2\overline{A}_{n\nu}} - 1)}$$

and

$$\alpha_\gamma = \ln\left|\frac{F}{\overline{P}}\right| = \ln\left[\frac{\sqrt{(e^{2\bar{a}_{n\nu}} - 1)}}{e^{\bar{a}_{n\nu}}}\right] + \tfrac{1}{2}\ln(e^{2\bar{a}_{n\nu} + 2\overline{A}_{n\nu}} - 1) - \tfrac{1}{2}\ln(e^{2\overline{A}_{n\nu}} - 1) \tag{21}$$

is the value to assign to the flat attenuation for the ideal transmission function in the specified stop band, except for an additive quantity of ln 2. The poles of F/P must be placed along the contour where

$$\alpha(s) = \ln\sqrt{\left(e^{2\bar{a}_{n\nu}+2\bar{A}_{n\nu}} - 1\right)}, \tag{22}$$

along certain equally-spaced lines of constant $\beta(s)$, as will be seen.

As shown by Klinkhamer a factor of ln 2 must be added to the above attenuations in the stop bands and a factor of ln 2 must be subtracted in the pass bands.[18] Thus, the following summarizes the requirements on the α/k in the various bands:

(a) In the pass bands:

$$\alpha_k \leq a_{nk} + \tfrac{1}{2}\ln(e^{2A_{nk}} - 1) - \ln 2;$$

(b) In the stop bands:

$$\alpha_k \geq \ln\left[\frac{\sqrt{(e^{2\bar{a}_{nk}}-1)}}{e^{\bar{a}_{nk}}}\right] + \tfrac{1}{2}\ln\left(e^{2\bar{a}_{nk}+2\bar{A}_{nk}}-1\right) - \tfrac{1}{2}\ln\left(e^{2A_{nk}}-1\right) + \ln 2, \text{ for } \bar{A}_{nk} \neq \infty,$$

$$\alpha_k \geq \bar{a}_{nk} + \bar{A}_{nk} - \tfrac{1}{2}\ln\left(e^{2\bar{A}_{nk}} - 1\right) + \ln 2, \text{ for } \bar{a}_{nk} > 2\text{neper},$$

$$\alpha_k \geq \bar{a}_{nk} + \ln 2, \text{ for } \bar{A}_{nk} = \infty.$$

Thus, for an elliptic low-pass filter the difference between flat loss in pass- and stop-band is

$$\alpha_{1n} = \bar{a}_{nk} + 2\ln 2 - \tfrac{1}{2}\ln(e^{2A_{nk}} - 1).$$

In db this is

$$\alpha_{1db} = \bar{a}_{db} - 10\log_{10}\left(10^{0.1A_{dbk}} - 1\right) + 12.04\text{db}.$$

If all the poles had been specified, the last constant would have been 6.02db and this would have yielded the usual correction factor for use with Rumpelt's template method. Note also that an interesting check results from using the material just after (2). As will be seen the total change in β in the upper passband must be $\frac{N\pi}{2}$. Hence,

$$\alpha_{1n} = \frac{K'}{K}\pi\frac{N}{2}, \tag{23}$$

so that

$$\bar{a}_{nk} = \tfrac{1}{2}\ln(e^{2A_{nk}} - 1) - 2\ln 2 + \frac{K'}{K}\frac{\pi}{2}N. \tag{24}$$

In db this is

$$\bar{a}_{dbk} = 10\log\left(10^{0.1A_{dbk}} - 1\right) + (N)\left(4.343\,\frac{K'}{K}\pi\right) - 12, \tag{25}$$

which is the formula first given by Darlington.[19] Further checks on the formulae for the α_k in the multi-band case have not yet been obtained. It appears that they are at least approximately valid.

The derivation of specifications on the w(s) for other network functions is not given, since the method should now be clear. Suffice it to note, as will shortly be discussed, that in the low-pass case, with ϵ constant $\phi(s) = \epsilon\cosh w(s)$, while a one-port resistance function is given by

$$R(s) = \frac{K_0}{\cosh w(s) + K_1}. \tag{26}$$

IV. DERIVATION OF RATIONAL FUNCTIONS WITH THE SPECIFIED EQUAL-RIPPLE APPROXIMATIONS TO THE SPECIFIED FLAT BANDS BY OPERATING ON w(s)

In the one low-band and in the one internal band cases the well-known method of obtaining equal ripple is to operate on the w(s) with the cosh function.[20,21,22,23] Thus, if filters are considered the characteristic function is found from

$$\phi(s) = \epsilon\cosh w(s). \tag{27}$$

(The w(s) may have been transformed, as done for the band-pass case in

Watanabe's article, using a special case of (10)). Previously Darlington and Cauer regarded w(s) as an image propagation function of a "reference filter," and this term still persists. It is possible that the physical idea may have delayed the discovery that $\frac{dw(s)}{ds}$ may have negative residues at its poles. The mathematical structure of w(s), however, clearly allows this possibility. Tuttle also finds that R(s) is given by (26), and finds a similar use for the cosh w(s) in the approximation of reactance and angle functions. Also, the function $R(s) = (1/\delta)\left[\tau_o - t(s)\right]$, where t(s) is a delay function, which was used by Ulbrich and Piloty in obtaining equal-ripple approximation to flat low-pass and band-pass delay, is easily found by operating on a w(s) with no poles on the $j\omega$-axis.[24] Useful as the functions found by use of the cosh have been in practice it is easily shown that for a flat band of α_k nepers

$$|\phi_K|_{max.} - |\phi_K|_{min.} = \epsilon(\cosh\alpha_k - \sinh\alpha_k)$$

and the mean value is

$$\tfrac{1}{2}\left[|\phi_k|_{max.} + |\phi_k|_{min.}\right] = \frac{\epsilon}{2}(\cosh\alpha_k + \sinh\alpha_k).$$

Hence, if ϵ has been fixed for the flat band with $\alpha_k=0$, and α_k in some band is given, there is no control over the ripple in that band. Thus, the cosh operator does not solve the general problem.

The tn, operating on, for example, the w(s) of (2) yields specified approximation in two bands, thus producing the $\phi(s)$ for elliptic low-pass filters. However, if the w(s) has additional flat bands one must simply accept the ripple that one obtains in these bands.

The dn operator, applied to a suitable w(s) with one internal band with non-zero flat α, yields the equal-ripple approximation of such an α by an odd function of frequency.[25]

In all the above one should note:

(1) the β in each flat band must be an integral multiple of π ($\arg[\ln\phi(s)]$ is a step approximation to β in the flat bands) and

(2) the operators cosh, tn and dn, operating on w(s), place zeros and poles (zeros only for the cosh) along contours of constant α where these contours are intersected by lines of constant β which are odd multiples of $\pi/2$.

Some research will show that this is what is done in general when finding equal-ripple approximations to flat bands. For example, the zeros of $\phi(s)$ for a low-pass Chebyshev filter with $R_1 \neq R_2$ lie on an ellipse.

With the aid of the potential analogy Klinkhamer has given a physically reasonable, though unproven, procedure for finding the zeros and poles of $\phi(s)$ from a knowledge of the contours of constant α and β of the non-pole w(s). It is assumed that the coefficients of the numerator polynomials in the integrands have been re-calculated, using equations based on changes in β, so that β changes by a multiple of π over each band except, possibly, the first low-pass band where, from $\omega=0$ to the first band edge, β may change by an odd multiple of $\pi/2$. This case will be called case (b), while the other will be called case (a). Then, with the w(s) properly chosen according to these principles and those in section V, one proceeds as follows in case (a).

Calculate β along all bands with constant α in, say, the upper half of the s plane. If $\alpha=0$ place a simple zero of $\phi(s)$ wherever $\beta=k\frac{\pi}{2}$ (k odd). If the flat transducer loss was specified as a_ν nepers in the pass band, place a simple zero of $\phi(s)$ on the contour $\alpha(s)=\ln\sqrt{e^{2a_\nu}-1}$ wherever $\beta=k\frac{\pi}{2}$ (k odd). Along a stop band with a maximum attenuation of infinity place a simple pole of $\phi(s)$ where $\beta=k\frac{\pi}{2}$ (k odd). If the maximum specified attenuation was not infinite place a simple pole of $\phi(s)$ on the contour $\alpha(s) = \ln\sqrt{e^{2a_{n\nu}+2A_{n\nu}}-1}$, wherever $\beta=k\frac{\pi}{2}$ (k odd). In case (b), use $\beta=k\pi$ (k=0,1,2,....). Note that w(s) may be Watanabe's generalized ideal transmission function, with flat bands along the real axis.

A proof that this method yields a solution to the general problem, or even a good first-order approximation thereto, has not yet been found. Physical considerations make it seem at least reasonable. Calculations on specific examples would provide valuable evidence. In the light of the above work the following unproved theorem is suggested:

The operators which must be applied to w(s) to yield the desired equal-ripple rational approximations with specified ripples in each band and specified fixed poles are, except for the possible addition of constants and/or taking reciprocals, inverses of Abelian integrals.

V. MAXIMALLY-FLAT APPROXIMATION OF THE FLAT BANDS

The most general low-pass $\phi(s)$ is easily derived as a limiting case of the theory above. From Watanabe's equation (14), for the pass band in $\left[-j\omega_p, j\omega_p\right]$,

$$\lim_{\omega_p \to 0} e^{w_\infty} = \lim_{\omega_p \to 0} \frac{1}{\omega_p^{2\gamma_\infty}}\left[\sqrt{s^2+\omega_p^2}+p\right]^{2\gamma_\infty} \longrightarrow \frac{s^{2\gamma_\infty}2^{2\gamma_\infty}}{\omega_p^{2\gamma_\infty}}\,, \tag{28}$$

so that

$$\lim_{\omega_p \to 0} \cosh w_\infty \longrightarrow \frac{s^{2\gamma_\infty}2^{2\gamma_\infty}}{2\omega_p^{2\gamma_\infty}}\,. \tag{29}$$

For complex poles,

$$\lim_{\omega_p \to 0} \cosh w_i \longrightarrow \tfrac{1}{2}\left\{\frac{2^{2\gamma_i}s_i^{2\gamma_i}}{\omega_p^{2\gamma_i}}\left[\frac{s^2}{p_i^2-p^2}\right]^{\gamma_i} \times \frac{2^{2\gamma_i}s_i^{*2\gamma_i}}{\omega_p^{2\gamma_i}}\left[\frac{s^2}{s_i^{*2}-s^2}\right]^{\gamma_i}\right\}. \tag{30}$$

Using a similar limiting process for real and imaginary poles, the generalized ϕ for the maximally-flat low-pass becomes

$$\phi = \frac{\sqrt{e^{2A_{pc}}-1}}{2}\left[\frac{2^{\gamma_\infty}}{\omega_p^{2\gamma_\infty}}\cdot s^{2\gamma_\infty}\cdot\prod_{\nu=1}^{h}\frac{2^{2\gamma_\nu}v_\nu^{2\gamma_\nu}}{\omega_p^{2\gamma_\nu}}\,\frac{s^{2\gamma_\nu}}{(s^2+v_\nu^2)^{\gamma_\nu}}\times\prod_{\mu=1}^{l}\frac{2^{2\gamma_\mu}}{\omega_p^{2\gamma_\mu}}\cdot(-u_\mu^2)^{\gamma_\mu}\,\frac{s^{2\gamma_\mu}}{(s^2-u_\mu^2)^{\gamma_\mu}}\right.$$
$$\left.\times\prod_{i=1}^{g}\frac{2^{4\gamma_i}(s_i s_i^*)^{2\gamma_i}}{\omega_p^{2\gamma_i}}\,\frac{s^{4\gamma_i}}{\left[(s^2-s_i^2)(s^2-s_i^{*2})\right]^{\gamma_i}}\right]. \tag{31}$$

ϕ will remain finite if

$$\lim_{\substack{\omega_p \to 0\\ A_{pc} \to 0}} \frac{\sqrt{e^{A_{pc}}-1}}{2}\left[\frac{2^{2\gamma_\infty}}{\omega_p^{2\gamma_\infty}}\cdot\prod_{\nu=1}^{h}\frac{2^{2\gamma_\nu}}{\omega_p^{2\gamma_\nu}}\cdot\prod_{\mu=1}^{l}\frac{2^{2\gamma_\mu}}{\omega_p^{2\gamma_\mu}}\prod_{i=1}^{g}\frac{2^{4\gamma_i}}{\omega_p^{2\gamma_i}}\right] = K = \text{constant.} \tag{32}$$

If $\phi(j\omega_p) = \sqrt{e^{2A_p}-1}$,

$$\phi = \sqrt{e^{2A_p}-1}\left(\frac{s}{\omega_p}\right)^{2\gamma_\infty}\cdot\prod_{\nu=1}^{h}\left(\frac{s}{\omega_p}\right)^{2\gamma_\nu}\left(\frac{\omega_p^2-v_\nu^2}{s^2+v_\nu^2}\right)^{\gamma_\nu}\cdot\prod_{\mu=1}^{l}\left(\frac{s}{\omega_p}\right)^{2\gamma_\mu}\left(\frac{\omega_p^2+u_\mu^2}{s^2-u_\mu^2}\right)^{\gamma_\mu}$$
$$\cdot\prod_{i=1}^{g}\left(\frac{s}{\omega_p}\right)^{4\gamma_i}\left[\frac{(\omega_p^2+s_i^2)(\omega_p^2+s_i^{*2})}{(s^2-s_i^2)(s^2-s_i^{*2})}\right]^{\gamma_i}. \tag{33}$$

With $l=g=0$, $2\gamma_\infty = 1$, $\gamma_\nu = 1$ and $\omega_p = 1$, the ϕ for symmetric low-pass filters follows as

$$\phi = \sqrt{e^{2A_p}-1}\; s^{2k+1}\prod_{\nu=1}^{h}\frac{1-v_\nu^2}{s^2+v_\nu^2}\,, \tag{34}$$

and with $2\gamma_\infty = 0$ and $\gamma_\mu = 1$, the ϕ for antimetric low-pass filters becomes

$$\phi = \sqrt{e^{2A_p}-1}\; s^{2h}\prod_{\nu=1}^{h}\frac{1-v_\nu^2}{s^2+v_\nu^2}\,. \tag{35}$$

The general ϕ for symmetrical, maximally-flat, band-pass filters follows upon substituting $s^2=p^2+1$ into (33). Then with $\omega_p^2=\omega_{pb}^2-1$, $V_\nu^2=v_\nu^2+1$, $-V_\mu^2=u_\mu^2-1$, $U_\mu^2=u_\mu^2-1$ and $p_i^2=s_i^2-1$,

$$\phi_{bp}=\sqrt{e^{2A_p}-1}\left(\frac{p^2+1}{\omega_{pb}^2-1}\right)^{\gamma_\infty}\cdot\prod_{\nu=1}^{h}\left(\frac{p^2+1}{\omega_{pb}^2-1}\right)^{\gamma_\nu}\left(\frac{\omega_{pb}^2-V_\nu^2}{p^2+V_\nu^2}\right)^{\gamma_\nu}\cdot\prod_{\mu=0}^{m}\left(\frac{p^2+1}{\omega_{pb}^2-1}\right)^{\gamma_\mu}\left(\frac{\omega_{pb}^2-V_\mu^2}{p^2+V_\mu^2}\right)^{\gamma_\mu}$$

$$\cdot\prod_{\mu=1}^{l-m-1}\left(\frac{p^2+1}{\omega_{pb}^2-1}\right)^{\gamma_\mu}\left(\frac{\omega_{pb}^2+U_\mu^2}{p^2-U_\mu^2}\right)^{\gamma_\mu}\cdot\prod_{i=1}^{g}\left(\frac{p^2+1}{\omega_{pb}^2-1}\right)^{2\gamma_i}\left[\frac{(\omega_{pb}^2+p_i^2)(\omega_{pb}^2+p_i^{*2})}{(p^2-p_i^2)(p^2-p_i^{*2})}\right]^{\gamma_i}\cdot \qquad (36)$$

With $l-m-1=g=0$, $\gamma_\infty=\gamma_0=\frac{1}{2}$, $\gamma_\mu=1$ $(\mu=1,2,\ldots m)$, $\gamma_\nu=1$ $(\nu=1,2,\ldots h)$, $V_0=0$, and $m+h=n$, one obtains

$$\phi=\pm\sqrt{e^{2A_p}-1}\,\frac{\omega_{pb}}{s}\left(\frac{p^2+1}{\omega_{pb}^2-1}\right)^{n+1}\prod_{\nu=1}^{n}\left(\frac{\omega_{pb}^2-V_\nu^2}{p^2+V_\nu^2}\right) \qquad (37)$$

for symmetric band-pass filters. With $\gamma_\infty=\gamma_0=0$, $\gamma_\mu=1$ $(\mu=1,2,\ldots m)$, $\gamma_\nu=1$ $(\nu=1,2,\ldots h)$, and $m+h=n$,

$$\phi=\sqrt{e^{2A_p}-1}\left(\frac{p^2+1}{\omega_{pb}^2-1}\right)^{n}\prod_{\nu=1}^{n}\left(\frac{\omega_{pb}^2-V_\nu^2}{p^2+V_\nu^2}\right) \qquad (38)$$

for antimetric band-pass filters. Equations (37) and (38) have been derived by Temes, using the reference filter concept.[26]

The form of the most general ϕ which approximates an arbitrary number of specified flat bands in the Taylor (maximally-flat) sense is easily deduced from the above. Thus, there is a multiple zero in each band with $\alpha=0$, complex zeros opposite each side of each pass band with $a_k\neq 0$, and complex poles opposite each side of each flat stop-band with $\bar{a}_k\neq\infty$. The specified poles are also exhibited, of course. For resistance and other network functions appropriate simple changes in form are introduced.[27] Thus, the maximally-flat approximations arise as limiting cases of the general theory.

ACKNOWLEDGEMENTS

The author is indebted to Dr. G. J. van der Maas and Claude Germain of the Northern Electric Co. Research and Development Laboratories for discussions and encouragement.

REFERENCES

1. Hitoshi Watanabe, "Approximation Theory for Filter Networks," IRE Transactions on Circuit Theory, Vol. CT-9, No.3, September, 1961, pp. 341-356.
2. Wilhelm Cauer, Theory of Linear Communication Networks, Vol. I (McGraw-Hill Book Company, Inc., Toronto, 1958).
3. J.F. Klinkhamer, "Empirical Determination of Wave-Filter Transfer Functions with Specified Properties," Philips Research Reports 3, 1948, Pt. 1, pp. 378-400.
4. V.V. Golubev, Lectures on Integration of the Equations of Motion of a Rigid Body About a Fixed Point (State Publishing House of Theoretical Technical Literature, Moscow, 1953), Chapter V.
5. Konrad Knopp, Theory of Functions, Vol. II (Dover Publications, New York, 1947), Chapter VI.
6. George Springer, Introduction to Riemann Surfaces (Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1957), Chapter X,

pp. 252-257 and pp. 292-298.

7. Golubev, *loc. cit.*
8. Einar Hille, *Analytic Function Theory*, Vol. II (Ginn and Company, Boston, 1962), Chapter XII.
9. Watanabe, *op. cit.*, pp. 342-343.
10. Tokuju Iedokoro, Tohoru Tsuchiya and Hitoshi Watanabe, "A New Calculation Method for the Design of Filters by Digital Computers with the Special Consideration of the Accuracy Problem," *IEEE International Convention Record*, Part 2, Circuit Theory, 1963, pp. 100-112.
11. Byron J. Bennett, "Synthesis of Electric Filters with Arbitrary Phase Characteristics," *IRE National Convention Record*, Pt. 5, Circuit Theory, 1953, pp. 19-26.
12. G. Szentirmai, "Theoretical Basis of a Digital Computer Program Package for Filter Synthesis," *Proceedings of the First Annual Allerton Conference on Circuit Theory and System Theory*, The University of Illinois, 1963.
13. James D. Schoeffler, "A Solution to the Approximation and Realization Problems for Crystal Ladder Filters," 1964, *IEEE International Convention Record*, Part I (March 23-26), pp. 282-288.
14. J.A.C. Bingham, "A New Method of Solving the Accuracy Problem in Filter Design," *IEEE Transactions on Circuit Theory*, CT-11, September, 1964, pp. 327-341.
15. R. Saal and E. Ulbrich, "On the Design of Filters by Synthesis," *IRE Transactions on Circuit Theory*, Vol. CT-5, No. 4 (December, 1958), pp. 297 and 309-313.
16. E. Rumpelt, "Schablonenverfahren für den Entwurf elektrischer Wellenfilter auf der Grundlage der Wellenparameter," *Telegraphen-Fernsprech- Funk- und Fernseh-Technik*, 31 Jahrgang, Heft 8, August 1942, pp. 203-210.
17. ________, "Methods of Numerical Filter Design," Part IV, *The Telecommunications Journal of Australia*, June, 1960, p. 273.
18. Klinkhamer, *loc. cit.*
19. Sidney Darlington, "Synthesis of Reactance 4-Poles which Produce Prescribed Insertion Loss Characteristics," *Journal of Mathematics and Physics*, September, 1939, pp. 331-332.
20. *Ibid.*, pp. 313-314.
21. Cauer, *op. cit.*, pp. 548-553.
22. Watanabe, *loc. cit.*
23. D.F. Tuttle, Jr., *Network Synthesis*, Vol. I, (John Wiley & Sons, New York, 1958), pp. 844-960.
24. Egbert Ulbrich and Hans Piloty, "Über den Entwurf von Allpässen, Tiefpässen und Bandpässen mit einer im Tschebyscheffsen Sinne approximierten konstanten Gruppenlaufzeit," *Archiv Elektrischen Übertragung*, Heft 10, 1960, pp. 451-456.
25. Sidney Darlington, "Realization of a Constant Phase Difference," *Bell System Technical Journal*, Vol. 29, January, 1950, pp. 94-104.
26. G.C. Temes, "The Synthesis of Filters with Maximally-flat Passband and Arbitrary Stopband Attenuation," *Canadian Electronics Engineering*, Vol. 8, No. 2, February, 1964, pp. 29-37.
27. Tuttle, *loc. cit.*

CONSIDERATIONS FOR THE DESIGN OF CRYSTAL FILTERS

E. CHRISTIAN, ITT Telecom and N.C. State Univ., Raleigh, N.C.
E. EISENMANN, Telefunken G.mb.H, Backnang, Germany

Abstract: The well established design methods for crystal filters are discussed and compared. Emphasis is placed on predicting the feasibility and the limitation of standard crystal filter designs.

Introduction: The practical realization of frequency selective networks makes it very often necessary to incorporate piezoelectric resonators in the circuit. These resonators, preferably quartz crystals, are used because they represent high-quality resonators of small or reasonable size. To include such elements in a circuit imposes restraints on the design due to three-element structure, the magnitude of the parameters and their influence on other components. The design of networks of this type postulates therefore:

(a) to contemplate circuits which are potentially capable of employing piezoelectric resonators, and
(b) to select related characteristic functions K(s) or transfer-loss functions H(s) such that the performance meets specification

It is frequently not feasible or not practical to substitute piezoelectric resonators in the resulting circuit. To precede a rather complicated circuit calculation by relating the feasibility to the primary design parameters (passband ripple, stopband attenuation and cut-off rate) is therefore economical. Such feasibility studies are possible for some of the typical bandpass realizations of Fig. 1 by

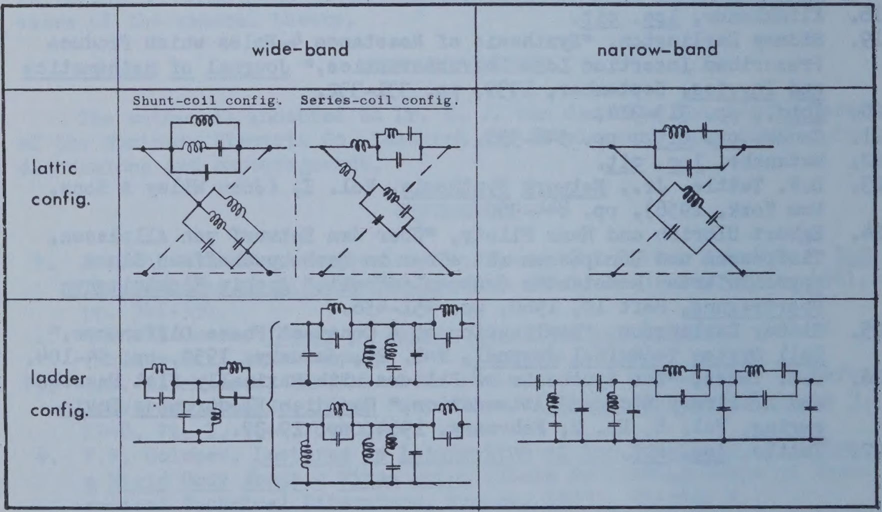


Fig. 1

considering the effect of the pole and zero pattern of the characteristic function K(s) and the transfer-loss function H(s) [SA-1]

$$K(s) = \frac{F(s)}{P(s)} \; ; \quad H(s) = \frac{E(s)}{P(s)}$$

$$\text{Operating Loss (db)} = \begin{cases} 10 \log_{10} \left(1+|K(s)|^2\right)_{s=jw} \\ \\ 10 \log_{10} |H(s)|^2_{s=jw} \end{cases} \qquad (1)$$

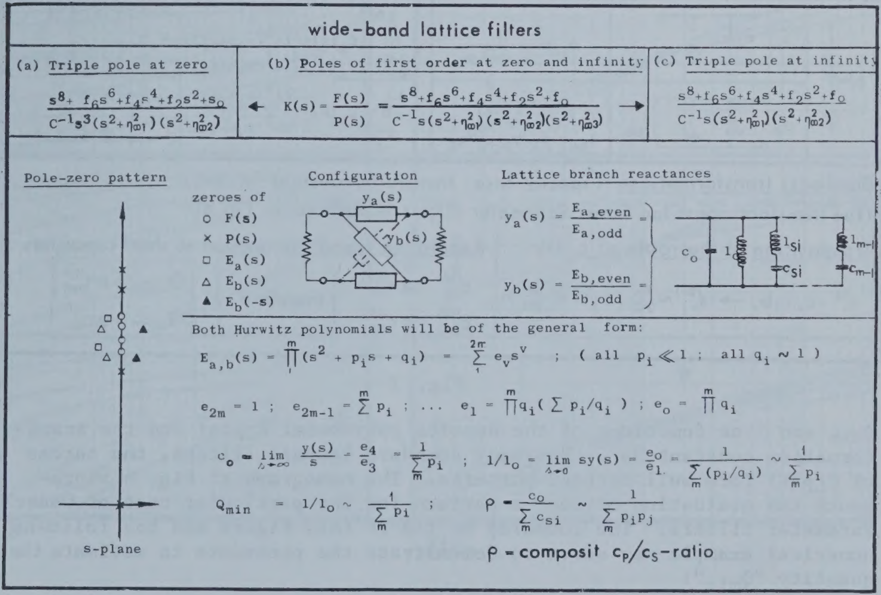
E(s) = Hurwitz polynomial ; P(s) = even or odd polynomial

F(s) = Polynomial with real coefficient which must be either even or odd in case of symmetrical or antimetrical filters

A distribution of almost all poles and zeroes of these functions about the center frequency is significant for many crystal filters. Their relative location within this area is very closely proportional to the relative bandwidth. This fact is used in many approximate solutions for the design.

1. Wideband Lattice Filters

The structure of an 8th degree characteristic function K(s) and some general design aspects for shunt-coil configurations are shown in Fig. 2. It is easy to prove that the shunt coils or the shunt capacitors of the two lattice branches become equal whenever a pole of third or higher order exists at zero or infinity, respectively. Analogous


wide-band lattice filters
(a) Triple pole at zero
(b) Poles of first order at zero and infinity
(c) Triple pole at infinity
Pole-zero pattern
s-plane
zeroes of
○ F(s)
× P(s)
□ $E_a(s)$
△ $E_b(s)$
▲ $E_b(-s)$
Configuration
$y_a(s)$
$y_b(s)$
Lattice branch reactances
c_o
l_o
l_{si}
c_{si}
l_{m-1}
c_{m-1}

$$\frac{s^8+f_6s^6+f_4s^4+f_2s^2+s_o}{C^{-1}s^3(s^2+\eta_{\infty1}^2)(s^2+\eta_{\infty2}^2)} \leftarrow K(s)=\frac{F(s)}{P(s)}=\frac{s^8+f_6s^6+f_4s^4+f_2s^2+f_o}{C^{-1}s(s^2+\eta_{\infty1}^2)(s^2+\eta_{\infty2}^2)(s^2+\eta_{\infty3}^2)} \rightarrow \frac{s^8+f_6s^6+f_4s^4+f_2s^2+f_o}{C^{-1}s(s^2+\eta_{\infty1}^2)(s^2+\eta_{\infty2}^2)}$$

$$y_a(s) = \frac{E_{a,even}}{E_{a,odd}} = \quad ; \quad y_b(s) = \frac{E_{b,even}}{E_{b,odd}} =$$

Both Hurwitz polynomials will be of the general form:

$$E_{a,b}(s) = \prod_{1}^{m}(s^2+p_is+q_i) = \sum_{1}^{2m} e_v s^v \quad ; \quad (\text{ all } p_i \ll 1 \; ; \text{ all } q_i \sim 1)$$

$$e_{2m} = 1 \; ; \quad e_{2m-1} = \sum^{m} p_i \; ; \; \ldots \; e_1 = \prod^{m} q_i \left(\sum p_i/q_i \right) \; ; \; e_o = \prod^{m} q_i$$

$$c_o = \lim_{s\to\infty} \frac{y(s)}{s} = \frac{e_4}{e_3} = \frac{1}{\sum_m p_i} \quad ; \quad 1/l_o = \lim_{s\to 0} sy(s) = \frac{e_o}{e_1} = \frac{1}{\sum_m (p_i/q_i)} \sim \frac{1}{\sum_m p_i}$$

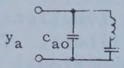
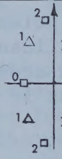
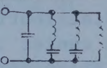
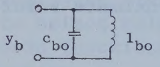
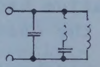
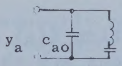
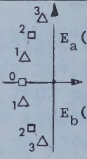
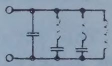
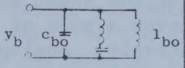
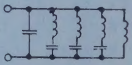
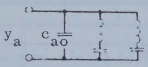
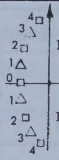
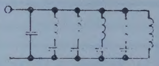
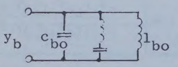
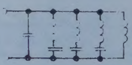
$$Q_{min} = 1/l_o \sim \frac{1}{\sum_m p_i} \quad ; \quad \rho = \frac{c_o}{\sum_m c_{si}} \sim \frac{1}{\sum_m p_i p_j}$$

ρ = composit c_p/c_s-ratio

Fig. 2

relations can also be drawn up for the dual series-coil configuration. Limiting factors for either type may be either the Q_{min} of the associated coils or ρ , the composite c_p/c_s-ratio of all crystals. To reduce the quantity "Q_{min}", it is necessary to increase the sum of the real parts of the Hurwitz polynomial E(s). This can be achieved by increasing the complexity (higher degree), by a decrease of the ripple or a combination of both. The composite c_p/c_s ratio on the other hand is improved by moving the zeroes of E(s) closer to the imaginary axis.

An estimate of the limiting quantities Q_{min} and ρ is possible in case of frequency-symmetrical specifications. A major part of the calculations, including the normalized reference circuit, can then be performed in the s-plane of a reference lowpass. By well-known lowpass-bandpass transformations, (Fig. 3), it is possible to express the quantities

DEGR.	LP-CONFIGUR.		$E_{LP}(s)$		BP-CONFIGUR.
N=5		$c_{ao}=\frac{1}{a_o+p_2}$		$E_a(s)=\begin{cases}(s+a_o)\\(s^2+p_2s+q_2)\end{cases}$	
		$c_{bo}=\frac{1}{p_1}$ $l_{bo}=p_1/q_1$		$E_b(s)=(s^2+p_1s+q_1)$	
N=7		$c_{ao}=\frac{1}{a_o+p_2}$		$E_a(s)=\begin{cases}(s+a_o)\\(s^2+p_2s+q_2)\end{cases}$	
		$c_{bo}=\frac{1}{p_1+p_3}$ $l_{bo}=p_1/q_1+p_3/q_3$		$E_b(s)=\begin{cases}(s^2+p_1s+q_1)\\(s^2+p_3s+q_3)\end{cases}$	
N=9		$c_{ao}=\frac{1}{a_o+p_2+p_4}$		$E_a(s)=\begin{cases}(s+a_o)\\(s^2+p_2s+q_2)\\(s^2+p_4s+q_4)\end{cases}$	
		$c_{bo}=\frac{1}{p_1+p_3}$ $l_{bo}=p_1/q_1+p_3/q_3$		$E_b(s)=\begin{cases}(s^2+p_1s+q_1)\\(s^2+p_3s+q_3)\end{cases}$	

Bandpass transformations (special case: transform. constant "a" ≫1):
(The transform.const. for Cauer Parameter Filters is defined in Fig. 4)

Trasposition of the roots of $E_{LP}(s)$

$$s_v^{(lp)}=a_v+jb_v \longrightarrow s_v^{(BP)}\sim\frac{a_v}{2a}\pm j\left(1+\frac{b_v}{2a}\right);$$

Reactance Transform. applied to shunt capacitors

$$\left.\begin{aligned}c_{ao}^{(BP)}&=a\,c_{ao}^{(LP)}\\c_{bo}^{(BP)}&=a\,c_{bo}^{(LP)}\end{aligned}\right\}\text{therefore}\left\{\boxed{\begin{aligned}Q_{a,min}&=a\,c_{ao}^{(LP)}\\Q_{b,min}&=a\,c_{bo}^{(LP)}\end{aligned}}\right.$$

Fig. 3

Q_{min} and ρ as functions of the Hurwitz polynomial $E_{LP}(s)$ and the transformation constant "a". For many standard lowpass filters, the zeroes of $E_{LP}(s)$ form well defined patterns. The nomograph of Fig. 4 represents the evaluation of such a pattern for the particular case of Cauer Parameter filters. The formulas on top of this Figure and the following numerical example may serve to demonstrate the procedure to estimate the quantity "Q_{min}":

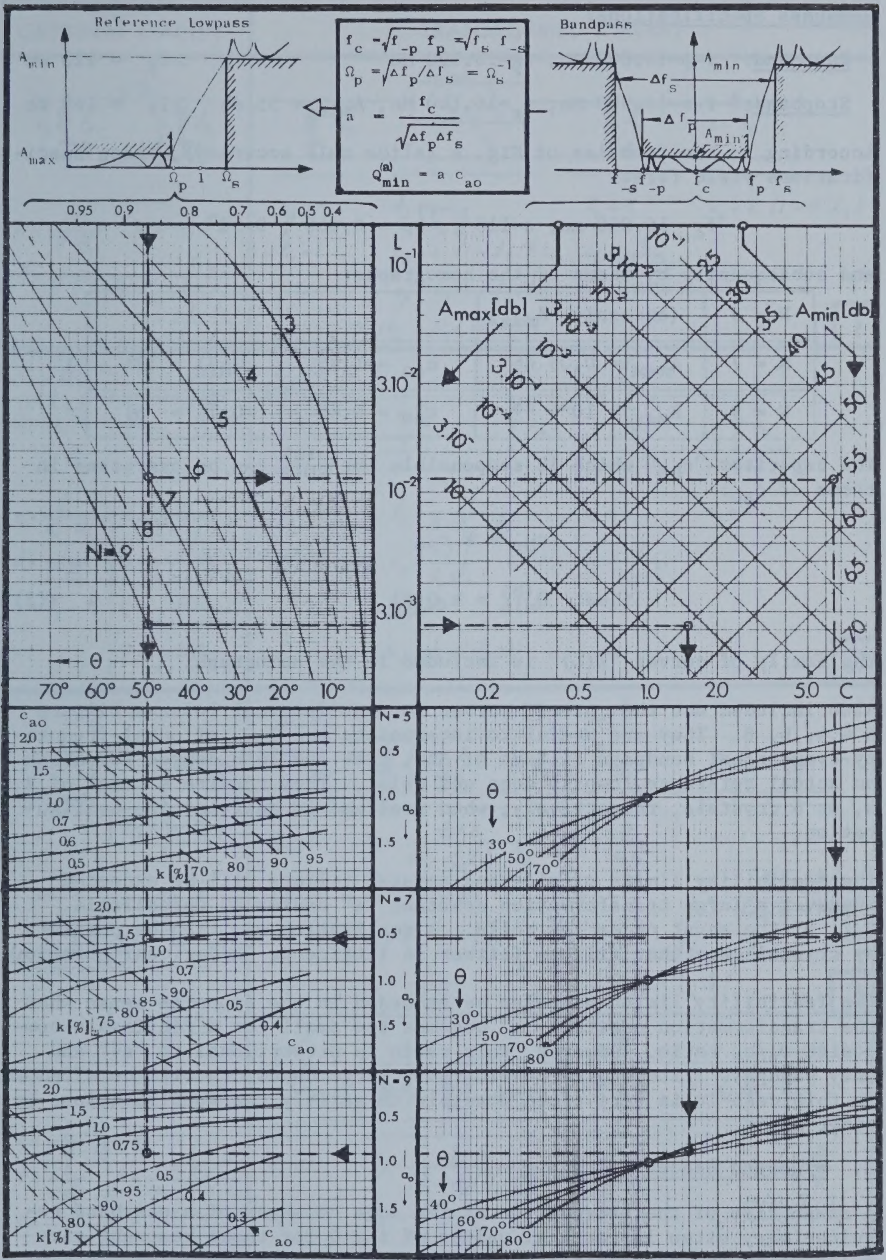

Reference Lowpass
A_min
A_max
Ω_p
1
Ω_s
f_c = √(f_-p f_p) = √(f_s f_-s)
Ω_p = √(Δf_p/Δf_s) = Ω_s^-1
a = f_c / √(Δf_p Δf_s)
Q_min^(a) = a c_ao
Bandpass
A_min
Δf_s
Δf_p
A_min
f_-s
f_-p
f_c
f_p
f_s
0.95
0.9
0.8
0.7
0.6
0.5
0.4
3
4
5
6
7
8
N=9
θ
70°
60°
50°
40°
30°
20°
10°
L
10^-1
3.10^-2
10^-2
3.10^-3
A_max[db]
A_min[db]
10^-1
3.10^-2
10^-2
3.10^-3
10^-3
3.10^-4
10^-4
3.10^-5
10^-5
25
30
35
40
45
50
55
60
65
70
0.2
0.5
1.0
2.0
5.0
C
c_ao
2.0
1.5
1.0
0.7
0.6
0.5
k[%]
70
80
90
95
N=5
0.5
1.0
1.5
a_o
θ
30°
50°
70°
N=7
2.0
1.5
1.0
0.7
0.5
0.4
90
85
80
75
k[%]
c_ao
θ
30°
50°
70°
80°
N=9
2.0
1.5
1.0
0.75
0.5
0.4
0.3
95
90
80
k[%]
c_ao
θ
40°
60°
70°
80°

Fig. 4

Bandpass Specifications:

Passband: f-p=16.000 Mc, f_p=16.120 Mc $\Delta f_p = 120$ kc

Stopband: f-s=15.980 Mc, f_s=16.140 Mc, $A_{min} = 55$ db $\Delta f_s = 160$ kc

According to the formulas of Fig. 4 (slide rule accuracy), these specifications yield first:

$$f_c \; 16.060 \text{ mc}; \; a=116 \; ; \; \Omega_p = 0.866 \; ; \; \theta=50^{\circ}$$

and subsequently by means of the nomographs:

N = 5	(excessive A_{max})		
N = 7	$A_{max} = 0.05$ db	$c_{ao} = 1.25$	$Q^{(a)}_{min} = 128$
N = 9	$A_{max} = 10^{-4}$ db	$c_{ao} = 0.60$	$Q^{(a)}_{min} = 69$

The capacitor "c_{bo}" which is responsible for $Q^{(b)}_{min}$ can be expressed in terms of c_{ao}:

$$c_{bo} = k \; c_{ao}$$

$$\text{thus:} \quad Q^{(b)}_{min} = k \; Q^{(a)}_{min} \tag{2}$$

The family of curves "k(%)" is included in the nomograph.

Also included are the cut-off curves for antimetrical lowpass filters N = 4, 6, 8. They are useful to interpolate the Q_{min} of quasi-frequency-symmetrical bandpass filters of 8th, 12th and 16th degree for which an actual reference lowpass does not exist. They require a total of 2, 4, or 6 crystals, respectively, when realized in half-lattice configuration.

The feasibility limits of wideband crystal filters will be exceeded whenever a large transformation constant "a" causes an unrealistic Q_{min} of the shunt or series coils. A realization may then be possible by crystal-capacitor filters (either in lattice or ladder configuration).

The feasibility limits may also be exceeded in the opposite sense when the transformation constant "a" becomes too small for a realizable composite c_p/c_s ratio. However, this ratio is proportional to "a^2" and only bandpass filters with considerably large bandwidth will be effected. In this case it is very often possible to satisfy the specifications by other means.

2. Wideband Ladder Filters

A steep rise of the attenuation at the edge or edges of a passband requires very often impractical values of the Q for those resonant circuits which are responsible for the closest attenuation poles. Piezoelectric resonators can be substituted for these critical circuits if the structure permits such a substitution. Network transformation similar to those shown in Fig. 5 may be employed to achieve such suitable structure. The transformation "4" of this Figure, originally suggested

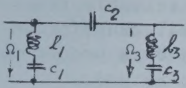
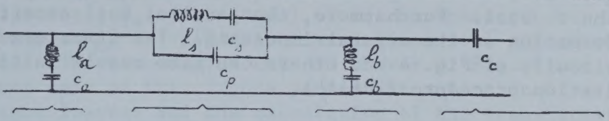
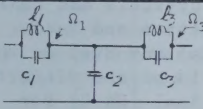
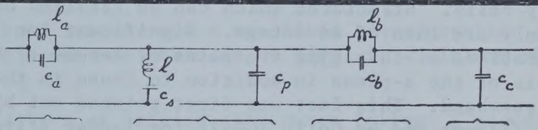
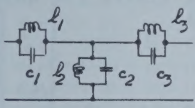
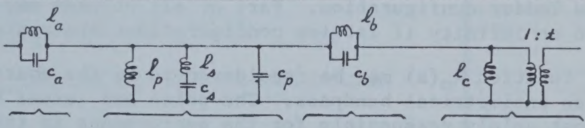
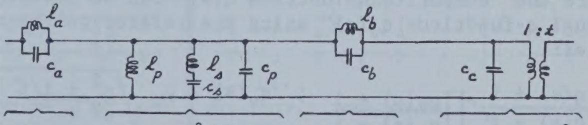
	ORIGINAL CIRCUIT	TRANSFORMED CIRCUIT
1	 c_2 Ω_1 ℓ_1 c_1 Ω_3 ℓ_3 c_3 $c = \frac{1}{\ell\,\Omega_1^2}$ $\ell = \frac{1}{c\,(\Omega_3^2-\Omega_1^2)}$ $A = c_1 - c$ $L = \frac{\ell\,\ell_1}{\ell-\ell_1}$ $M = \frac{\ell\,\ell_3}{\ell+\ell_3}$	 ℓ_a c_a ℓ_s c_s c_p ℓ_b c_b c_c $c_a = A$ $\ell_a = L$ $\ell_s = \frac{1}{c_s\Omega_1^2}$ $c_s = c\,(1+\ell_3/\ell)$ $c_p = c_2(1+\ell_3/\ell)$ $c_p/c_s = \Omega_1^2/(\Omega_3^2-\Omega_1^2)$ $\ell_b = M$ $c_b = \frac{1}{\ell_b\,\Omega_1^2}$ $c_c = c\,(1+\ell/\ell_3)$
2	 ℓ_1 Ω_1 ℓ_3 Ω_3 c_1 c_2 c_3 $c = \frac{\Omega_3^2}{\Omega_1^2-\Omega_3^2}c_2$ $\ell = \frac{1}{c\,\Omega_1^2}$ $L = \ell_1 - \ell$ $A = \frac{c\,c1}{c-c_1}$ $B = \frac{c\,c_2}{c+c_2+c_3}$ $C = \frac{c\,c_3}{c+c_2+c_3}$ $D = \frac{c_2c_3}{c+c_2+c_3}$	 ℓ_a c_a ℓ_s c_s c_p ℓ_b c_b c_c $c_a = A$ $\ell_a = L$ $\ell_s = \frac{1}{B\,(\Omega_1^2-\Omega_3^2)}$ $c_s = \frac{1}{\ell_s\,\Omega_3^2}$ $c_p = B$ $c_p/c_s = \Omega_3^2\,/(\Omega_1^2-\Omega_3^2)$ $c_b = C$ $\ell_b = \frac{1}{c_b\,\Omega_1^2}$ $c_c = D$
3	$\eta_1 = (\ell_1c_1)^{-\frac{1}{2}}$ $\eta_2 = (\ell_2c_2)^{-\frac{1}{2}}$ $\eta_3 = (\ell_3c_3)^{-\frac{1}{2}}$ $\eta_2<\eta_3<\eta_1$  ℓ_1 ℓ_3 c_1 ℓ_2 c_2 c_3	 ℓ_a c_a ℓ_p ℓ_s c_s c_p ℓ_b c_b $1:t$ ℓ_c $\ell_a = L$ $c_a = A$ $\ell_s = \frac{\eta_3^2/B}{(\eta_3^2-\eta_2^2)(\eta_1^2-\eta_3^2)}$ $\ell_p = \frac{tM}{t+DM\eta_1^2}$ $\left.\begin{array}{l} c_p = B + D/t \\ c_s = \frac{1}{\ell_s\eta_3^2} \end{array}\right\\} c_p/c_s = \frac{\eta_3^4\,[1+c_3/(c+c_2)]}{(\eta_3^2-\eta_2^2)(\eta_1^2-\eta_3^2)}$ $\ell_b = \frac{1}{c_b\,\eta_1^2}$ $c_b = C/t$ $t = 1 + c_2/c$ $\ell_c = \frac{t^2\ell_3/D\eta_1^2}{\ell_3\eta_2^2/\eta_1^2 - 1/D\eta_2^2}$
4	$\eta_1<\eta_3<\eta_2$ $c = \frac{\eta_3^2-\eta^2}{\eta_1^2-\eta_3^2}\;c_2$ $\ell = \frac{1}{c\,\eta_1^2}$ $A = \frac{c\,c_1}{c-c_1}$ $L = \ell_1 - \ell$ $B = \frac{c\;c_2}{c+c_2+c_3}$ $C = \frac{c\;c_3}{c+c_2+c_3}$ $D = \frac{c_2\,c_3}{c+c_2+c_3}$ $M = (\ell+\ell_2)[1+c_3/(c_2+c)]$	 ℓ_a c_a ℓ_p ℓ_s c_s c_p ℓ_b c_b $1:t$ c_c $\ell_a = L$ $c_a = A$ $\ell_s = \frac{\eta_3^2/B}{(\eta_3^2-\eta_2^2)(\eta_1^2-\eta_3^2)}$ $\ell_p = \frac{tM}{t+DM\eta_2^2}$ $\left.\begin{array}{l} c_p = B + \frac{D/\eta_2^2}{t\;\eta_1^2} \\ c_s = 1/\,\ell_s\eta_3^2 \end{array}\right\\} c_p/c_s = \frac{\eta_3^4[\,1+c_3\eta_2^2/(c\,\eta_1^2+c_2\eta_2^2)]}{(\eta_3^2-\eta_2^2)(\eta_1^2-\eta_3^2)}$ $\ell_b = \frac{1}{c_b\eta_1^2}$ $c_b = C/t$ $t = 1 + \frac{c_2\eta_2^2}{c\;\eta_1^2}$ $c_c = \frac{D\,(1-\eta_2^2/\eta_1^2)}{t^2}$

Fig. 5

by Poschenrieder ([PO-1]), and the transformation "3" are especially useful. The resulting structures prevent a signal flow at the pole frequency by a low impedance shunt at the highly stable series resonance of the crystal. Furthermore, the parallel coil permits a suitable transformation of the crystal impedance. The structure of all transformed circuits of Fig. 4 and others can also result initially by a proper realization procedure ([HA-1]).

It should also be mentioned that narrow bandstop filters can also be employed to supplement the performance of conventional LC-filters ([SA-2]).

3. Narrow-band Lattice and Ladder Filters

The performance of a conventional narrow bandpass filter will deteriorate considerably due to the inherent losses of actual components, especially coils. Structures which can be realized by crystals and capacitors only are then of advantage. Significant for lattice and ladder configurations of this type are pairs of zeroes of K(s) and H(s) on the real axis of the s-plane in addition to those in the immediate vicinity of the passband. This fact was first pointed out by Saal [SA-2] for lattice filters and by Colin [CO-1] for ladder filters. The pertinent function K(s) is then:

$$K(s) = \frac{(s^2 - a_o^2)}{(1 + a_o^2)s} K_o(s) = \frac{(s^2 - a_o^2)}{(1 + a_o^2)s} \underbrace{C \prod_{\nu} \frac{(s^2 + \eta_{o\nu}^2)}{(s^2 + \eta_{\infty\nu}^2)}}_{K_o(s)} \qquad (3)$$

All attenuation poles must occur at finite frequencies in case of genuine ladder configuration. Part or all of them may be transposed to zero or infinity if lattice configurations are contemplated.

The function $K_o(s)$ may be considered to be the characteristic function of an antimetrical bandpass. The poles and zeroes of this function are almost solely responsible for the performance in the passband and its vicinity. Cauer's q-functions are a convenient means to design a $K_o(s)$ which provides a Chebyshev variation within the passband limits for any arbitrary set of attenuation poles ([CA-1], pgs 548-560)

$$K_o(s) = \sqrt{10^{0.1\, A_{max}} - 1}\; \frac{q^2(s) + 1}{q^2(s) - 1} \qquad (4)$$

where the "composite q-function q(s)" can be calculated from the "individual q-functions $q_v(s)$" which are related to the specified attenuation poles:

$$\frac{q(s) + 1}{q(s) - 1} = \prod_{\nu} \frac{q_v(s) + 1}{q_v(s) - 1} \; ; \quad q_v(s) = m_v \sqrt{\frac{s^2 + 1/\tau}{s^2 + \tau}}$$

$$\tau = \sqrt{f_p/f_{-p}} \; ; \quad m_v = \sqrt{\frac{\eta_{\infty\nu}^2 - \tau}{\eta_{\infty\nu}^2 - 1/\tau}} \qquad (5)$$

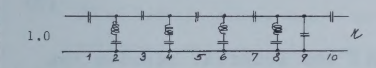
The antimetrical function $K_o(s)$ is then supplemented by the symmetry factor "$(s^2 - a_o^2)/(1 + a_o^2)s$" which provides for a first order pole each at zero and infinity. Its magnitude is close to unity in the immediate vicinity of the passband.

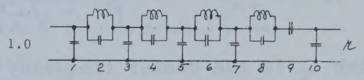
The value "a_o" is of significance for the composite c_p/c_s ratio of either branch. A rather accurate estimate of these quantities is possible in the case of frequency-symmetrical performances for which the pertinent Hurwitz polynomials $E_a(s)$ and $E_b(s)$ of the bandpass can be related to the polynomials $E_{LP}^{(a)}(s)$ $E_{LP}^{(b)}(s)$ of an antimetrical reference lowpass. (See Fig. 3). The even-degree cut-off curves of Fig. 4 and the formulas in the top part of this figure permit the selection of a suitable table reference lowpass and the calculation of the transformation constant "a". Expressed in terms of the real parts "a_v" of $E_{LP}^{(a)}(s)$ or $E_{LP}^{(b)}(s)$, the quantity becomes:

$$\rho = \frac{c_p}{\sum\limits_i c_{si}} = \frac{a}{(a_o + 1/a_o)} \frac{1}{\sum\limits_v a_v} \qquad (6)$$

For a selected reference lowpass and a specified transformation constant, this quantity becomes a maximum where $(a_o + 1/a_o)$ is a minimum which occurs for $a_o = 1$.

The well-known realization of lattice structures follows a procedure analogous to the one shown in Fig. 2. Crystal capacitor ladder structures, however, require special attention. Due to the factor $(s^2 - a_o^2)$ in the characteristic function, the open-and the short-circuit impedance





	z_{open}	z_{shunt}
c_1 =	1.73006	1.73008
ℓ_2 =	168.398	168.405
c_2 =	0.00595	0.00595
c_3 =	2.61987	2.61991
ℓ_4 =	17.8752	17.8781
c_4 =	0.05607	0.05606
c_5 =	26.1639	26.1552
ℓ_6 =	1.18146	1.18005
c_6 =	0.84835	0.84936
c_7 =	409.336	409.817
ℓ_8 =	0.10428	0.10431
c_8 =	9.61105	9.60876
c =	3820.21 $= c_9$	6612.72 $= c_9 + c_{10}$
r	0.0006196	

Open and shunt impedances:

$$z_{open} = \frac{e_7 s^6 + e_5 s^4 + e_3 s^2 + e_1}{2s^7 + (e_6 + f_6)s^5 + (e_4 + f_4)s^3 + (e_2 + f_2)s}$$

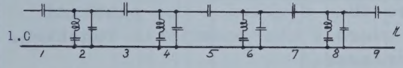
$$z_{shunt} = \frac{(e_6 - f_6)s^6 + (e_4 - f_4)s^4 + (e_2 - f_2)s^2 + (e_0 - f_0)}{e_7 s^7 + e_5 s^5 + e_3 s^3 + e_1 s}$$

$$F(s) = s^8 + f_6 s^6 + f_4 s^4 + f_2 s^2 + f_0$$

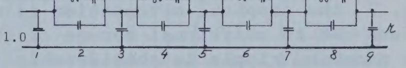
$$E(s) = s^8 + e_7 s^7 + e_6 s^6 + e_5 s^5 + e_4 s^4 + e_3 s^3 + e_2 s^2 + e_1 s + e_0$$

$f_0 = - e_0$ due to the zero on the real axis

z_{open}	z_{shunt}
0.576120	0.576120
0.005906	0.005906
168.9161	169.9161
0.38316	0.38316
0.05534	0.05534
18.0282	18.0282
0.038614	0.038614
0.836242	0.836242
1.193093	1.193093
0.002470	0.002470
9.40665	9.406659
0.106065	0.106065
0.000153 $= c_{9,10}$	0.000266 $= c_9$
1592	



$c_1 = 1.730$
$\ell_2 = 168.0$ $c_2 = 0.00595$ $\eta_2 = 0.998855$ $c_{p2} = 1.766$ $\rho = 297$
$c_3 = 0.854$
$\ell_4 = 168.4$ $c_4 = 0.00595$ $\eta_4 = 0.998855$ $c_{p4} = 1.488$ $\rho = 250$
$c_5 = 0.713$
$\ell_6 = 168.4$ $c_6 = 0.00595$ $\eta_6 = 0.998855$ $c_{p6} = 1.488$ $\rho = 250$
$c_7 = 0.854$
$\ell_8 = 168.4$ $c_8 = 0.00595$ $\eta_8 = 0.998855$ $c_{p8} = 1.766$ $\rho = 297$
$c_9 = 1.730$
$r = 1.000$



$c_1 = 0.576$
$\ell_2 = 480.5$ $c_2 = 0.00208$ $\eta_2 = 1.002905$ $c_{p2} = 0.592$ $\rho = 289$
$c_3 = 1.079$
$\ell_4 = 480.7$ $c_4 = 0.00208$ $\eta_4 = 1.00306$ $c_{p4} = 0.544$ $\rho = 261$
$c_5 = 1.189$
$\ell_6 = 480.7$ $c_6 = 0.00208$ $\eta_6 = 1.00306$ c_{p6} 0.544 $\rho = 261$
$c_7 = 1.079$
$\ell_8 = 480.5$ $c_8 = 0.00208$ $\eta_8 = 1.002905$ $c_{p8} = 0.592$ $\rho = 284$
$c_9 = 0.576$
$r = 1.000$

Fig. 6

at either terminal pair is one degree lower than the degree of the characteristic function. Consequently, the ladder development of either impedance will leave one capacitor undetermined. (See Fig. 6).

A convenient means to overcome any ambiguity of the last capacitors is to realize the circuit in shunt-and open-circuit condition. The comparison of the values of all other elements can be used as a criterion for adequate precision of the calculation. The numerical examples of Fig. 6 are related to a 10.7 Mc filter with all poles at 13.5 Kc above or below the passband, respectively. The left circuit was calculated with 24, the right circuit with 28 significant digits. As the comparison of "shunt" and "open" shows, 24 significant digits were just adequate.

Any realization of ladder structures will normally yield the configurations of crystal resonators only if the attenuation poles above and below the passband alternate. Any other sequence will require subsequent network transformation similar to those indicated at the bottom of Fig. 6. These particular transformations were performed to yield equal inductance for all the crystals of one structure.

The lowest c_p/c_s ratio within the final structure can also be estimated without an elaborate calculation by means of the following formula:

$$\rho = \sqrt{\frac{f_c}{\Delta f}} \qquad (7)$$

in which f_c is the center frequency and Δf the interval between the lowest (or highest) attenuation pole and the upper (or lower) passband limit. For the left hand circuit of Fig. 6, the lowest pole was at 10.685 Mc and the upper passband limit 10.703 Mc. Equation (7) yields then $\rho = 243$, sufficiently close to the actual value.

4. Lattice-Ladder Hybrid Structures

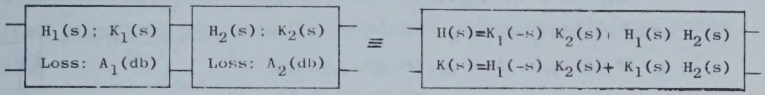
Lattice configurations are a special case of structures which provide more than one path for the transmission of the signal. Attenuation poles occur at those frequencies at which cancellation of the energy flow takes place. This is an inherent disadvantage whenever high stopband attenuations are specified. It is therefore often convenient to cascade several sections in tandem and to employ lattice configurations only whenever this particular structure offers advantages, for instance better c_p/c_s ratio. Three different approaches are, at present, in use.

(a) Cascading of identical lattice configurations. According to Fig. 7, the stopband attenuation is more than doubled. The passband widens due to additional zeroes of the overall characteristic function. The passband ripple is considerably increased.

(b) Bisection of a symmetrical network. The priority to the solution goes to H. Piloty [PT - 1]. A different solution, developed by H. Adler and E. H. Bradley ([GU - 1], pg. 467), was recently applied by Szentirmai [SZ - 1] to lattice configurations. According to Piloty's method, it is first necessary to augment the denominator polynomial of an overall characteristic function K(s) by factors which make it a perfect square. The more factors required to achieve this the less economical will be the overall structure. A realization of a network by bisection is therefore recommended only in case of coinciding attenuation poles.

(c) Lattice transformations applied to suitable sections of an established ladder structure. This solution was first indicated by Szentirmai [SZ - 1] and later supplemented by J.A.C. Bingham [BI - 1]. A different approach is presented in Fig. 8. To achieve optimum conditions of the final circuit, it is useful to realize the initial ladder circuit with an appropriate sequence of pole frequencies. (See the numerical example).

CASCADED LATTICE FILTERS


$H_1(s)$; $K_1(s)$
Loss: A_1(db)
$H_2(s)$: $K_2(s)$
Loss: A_2(db)
≡
$H(s)=K_1(-s)\ K_2(s)+\ H_1(s)\ H_2(s)$
$K(s)=H_1(-s)\ K_2(s)+\ K_1(s)\ H_2(s)$

Special case : identical networks

$$K(s) = 2\ K_1(s)\ \frac{E_{1,odd}(s)}{P_1(s)}$$

$$H(s) = K_1(-s)\ K_1(s) + H_1^2(s)$$

Approximation of the performance:

passband : $|K(s)| \sim 2\,|K_1(s)| + |K_1(s)|^3$

Stopband : $A(db) \sim 2\ A_1(db) + 6\ db$

Numerical example : Crystal-capacitor bandpass of 6th degree

Center frequ. f_c = 456 kc; passband limits : f_c ±140 cps ; p = 25 %

attenuation poles: $f_c \pm 6$ kc ; A_{min} = 56 db

These specifications can be satisfied by a lattice related to the following characteristic function $K_1(s)$:

$$K_1(s)=\frac{F_1(s)}{P_1(s)}=\frac{(s^2+\Omega_{02}^2)(s^2+\Omega_{01}^2)(s^2-1)}{C^{-1}s(s^2+\Omega_{\infty1}^2)(s^2+\Omega_{\infty2}^2)}=\frac{s^6+1.000000154s^4-1.000000156s^2-0.999999998}{(3.00435s^5+6.010766s^3+3.004356s).10^{-3}}$$

Hurwitz Polynomial :

$$E_1(s) = s^6+2.001330026s^5+3.002656577s^4+4.002662126s^3+3.002656575s+$$
$$2.001330022s + 1.00000000$$

$$E_{1,odd}(s) = 2.001330026s^5+4.002662126s^3+2.001330022s$$

Overall characteristic function K(s) of the tandem connection :

$$K(s)=\frac{F_1(s)\ E_{1,odd}(s)}{P_1^2(s)}=\frac{(s^2+\Omega_{01}^2)(s^2+\Omega_{02}^2)(s^2-1)(2.001330026s^4+\ldots.\ 2.001330022)}{s(3.00435s^4+6.010766s^2+3.004356)^2 10^{-6}}$$

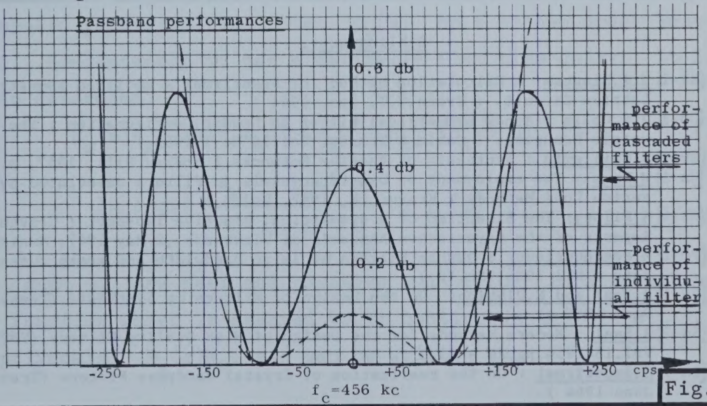
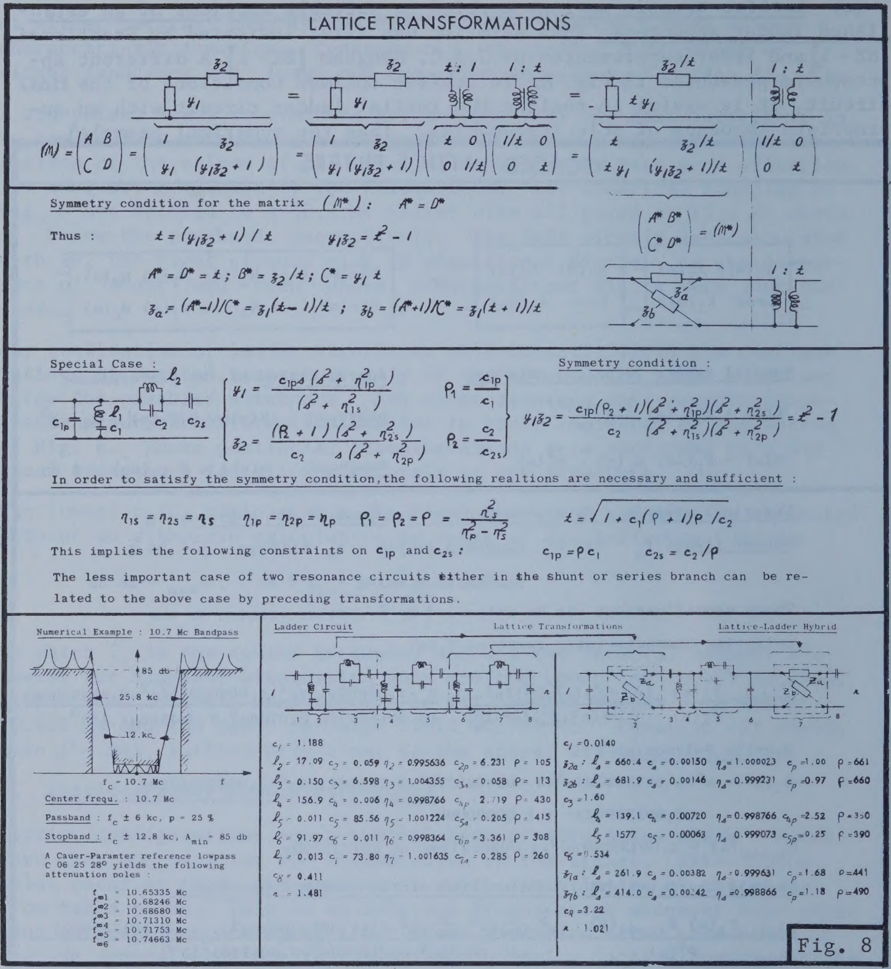


Fig. 7



LATTICE TRANSFORMATIONS

$$(M) = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & z_2 \\ y_1 & (y_1 z_2 + 1) \end{pmatrix} = \begin{pmatrix} 1 & z_2 \\ y_1 & (y_1 z_2 + 1) \end{pmatrix} \begin{pmatrix} t & 0 \\ 0 & 1/t \end{pmatrix} \begin{pmatrix} 1/t & 0 \\ 0 & t \end{pmatrix} = \underbrace{\begin{pmatrix} t & z_2/t \\ t\,y_1 & (y_1 z_2 + 1)/t \end{pmatrix}}_{\begin{pmatrix} A^* & B^* \\ C^* & D^* \end{pmatrix} = (M^*)} \begin{pmatrix} 1/t & 0 \\ 0 & t \end{pmatrix}$$

Symmetry condition for the matrix (M^*) : $A^* = D^*$

Thus : $t = (y_1 z_2 + 1)/t \qquad y_1 z_2 = t^2 - 1$

$$A^* = D^* = t\,;\; B^* = z_2/t\,;\; C^* = y_1 t$$

$$z_a = (A^* - 1)/C^* = z_1(t - 1)/t\;;\quad z_b = (A^* + 1)/C^* = z_1(t + 1)/t$$

Special Case :

$$\left\{\begin{aligned} y_1 &= \frac{c_{1p}s\,(s^2 + \eta_{1p}^2)}{(s^2 + \eta_{1s}^2)} & \rho_1 &= \frac{c_{1p}}{c_1} \\ z_2 &= \frac{(\rho_2 + 1)}{c_2}\,\frac{(s^2 + \eta_{2s}^2)}{s\,(s^2 + \eta_{2p}^2)} & \rho_2 &= \frac{c_2}{c_{2s}} \end{aligned}\right\}$$

Symmetry condition :

$$y_1 z_2 = \frac{c_{1p}(\rho_2 + 1)}{c_2}\,\frac{(s^2 + \eta_{1p}^2)(s^2 + \eta_{2s}^2)}{(s^2 + \eta_{1s}^2)(s^2 + \eta_{2p}^2)} = t^2 - 1$$

In order to satisfy the symmetry condition, the following realtions are necessary and sufficient :

$$\eta_{1s} = \eta_{2s} = \eta_s \qquad \eta_{1p} = \eta_{2p} = \eta_p \qquad \rho_1 = \rho_2 = \rho = \frac{\eta_s^2}{\eta_p^2 - \eta_s^2} \qquad t = \sqrt{1 + c_1(\rho + 1)\rho / c_2}$$

This implies the following constraints on c_{1p} and c_{2s} : $\quad c_{1p} = \rho\, c_1 \qquad c_{2s} = c_2/\rho$

The less important case of two resonance circuits either in the shunt or series branch can be related to the above case by preceding transformations.

Numerical Example : 10.7 Mc Bandpass

Center frequ. : 10.7 Mc

Passband : f_c ± 6 kc, p = 25 %

Stopband : f_c ± 12.8 kc, A_{min} = 85 db

A Cauer-Paramter reference lowpass C 06 25 28° yields the following attenuation poles :

$f_{\infty 1}$ = 10.65335 Mc
$f_{\infty 2}$ = 10.68246 Mc
$f_{\infty 3}$ = 10.68680 Mc
$f_{\infty 4}$ = 10.71310 Mc
$f_{\infty 5}$ = 10.71753 Mc
$f_{\infty 6}$ = 10.74663 Mc

Ladder Circuit

$c_1 = 1.188$
$\ell_2 = 17.09 \quad c_2 = 0.059 \quad \eta_2 = 0.995636 \quad c_{2p} = 6.231 \quad \rho = 105$
$\ell_3 = 0.150 \quad c_3 = 6.598 \quad \eta_3 = 1.004355 \quad c_{3s} = 0.058 \quad \rho = 113$
$\ell_4 = 156.9 \quad c_4 = 0.006 \quad \eta_4 = 0.998766 \quad c_{4p}\ 2.719 \quad \rho = 430$
$\ell_5 = 0.011 \quad c_5 = 85.56 \quad \eta_5 = 1.001224 \quad c_{5s} = 0.205 \quad \rho = 415$
$\ell_6 = 91.97 \quad c_6 = 0.011 \quad \eta_6 = 0.998364 \quad c_{6p} = 3.361 \quad \rho = 308$
$\ell_7 = 0.013 \quad c_7 = 73.80 \quad \eta_7 = 1.001635 \quad c_{7s} = 0.285 \quad \rho = 260$
$c_8 = 0.411$
$r = 1.481$

Lattice-Ladder Hybrid

$c_1 = 0.0140$
$z_{2a} : \ell_s = 660.4 \quad c_s = 0.00150 \quad \eta_s = 1.000023 \quad c_p = 1.00 \quad \rho = 661$
$z_{2b} : \ell_s = 681.9 \quad c_s = 0.00146 \quad \eta_s = 0.999231 \quad c_p = 0.97 \quad \rho = 660$
$c_3 = 1.60$
$\ell_4 = 139.1 \quad c_4 = 0.00720 \quad \eta_s = 0.998766 \quad c_{4p} = 2.52 \quad \rho = 350$
$\ell_5 = 1577 \quad c_5 = 0.00063 \quad \eta_s\ 0.999073 \quad c_{5p} = 0.25 \quad \rho = 390$
$c_6 = 0.534$
$z_{7a} : \ell_s = 261.9 \quad c_s = 0.00382 \quad \eta_s = 0.999631 \quad c_p = 1.68 \quad \rho = 441$
$z_{7b} : \ell_s = 414.0 \quad c_s = 0.00242 \quad \eta_s = 0.998866 \quad c_p = 1.18 \quad \rho = 490$
$c_8 = 3.22$
$r\ 1.021$

Fig. 8

Bibliography.

[BI-1] J.A.Bingham: A Calculation Method for Szentirmai's modified symmetrical Lattice Circuit. (Trans.C.T. June 1964)

[CO-1] J.E.Colin : A new symmetrical Filter in Ladder Configuration : The parametric Bandpass. (Cable & Transm. Vol.16 No.1 1962)

[GU-1] E.Guillemin : Synthesis of Passive Networks (McGraw Hill, 1957, p. 467)

[HA-1] W.Haas : The employment of Quartz crystals in Filters calculated by Inserion Loss Methods (Frequ.,16, 1962, Vol.5)

[PI-1] H.Piloty: The Bisection of a symmetrical reactive Two-port (Convention Reports of the Bayerischen Akademie d. Wissensch.,2 May 1947, Munich)

[PO-1] W.Poschenrieder : Quartz Filters of large Bandwidth and steep cut-off Rate (NTZ 1956, No.12)

[SA-1] R.Saal and E.Ulbrich: On the Design of Filters by Synthesis.(Trans.CT., Dec. 1958)

[SA-2] R.Saal : Modern Methods for the Design of High Quality Filters.(VDE Reports, Vol.18,1954 pg.29 - 33)

[SZ-1] G.Szentirmai : On the realization of Crystal Bandpass Filters (Trans.CT., June 1964)

AN APPROXIMATION TO A SPECIFIED MAGNITUDE AND MAXIMALLY FLAT DELAY USING NON-MINIMUM PHASE NETWORKS

S. KARNI AND NASIR AHMED
Department of Electrical Engineering
University of New Mexico, Albuquerque, New Mexico

SUMMARY

A procedure to approximate a maximally flat time delay and a specified magnitude response is developed using non-minimum phase networks.

INTRODUCTION

We consider a system represented by the block diagram given in Fig. 1(a). It is desired to modify the amplitude spectrum of the signal s(t) without phase distortion. Then a procedure to obtain a network with a specified magnitude characteristic and maximally flat delay is required. Such a network may then be introduced as shown in Fig. 1(b).

Let $F_k(s) = \frac{1}{D_k(s)}$ be an all-pole function which has a maximally flat time delay characteristic. Then it is clear that the transfer function

$$H_k(s) = \frac{N_k(s)\,N_k(-s)}{D_k^2(s)} \tag{1}$$

also has a maximally flat time delay since the zeros of the numerator of $H_k(s)$ possess quadrant symmetry. In what follows we develop a method to obtain $N_k(s)\,N_k(-s)$ such that $H_k(s)$ is realizable.

ANALYSIS

Consider the following problem, the analysis of which leads to a straightforward generalization. Let $D_k(s)$ be a second order Bessel polynomial. Consequently it has a maximally flat time delay. It is required to determine

$$N_k(s)N_k(-s) = a_k\omega^4 + b_k\omega^2 + c_k, \quad (c_k > 0) \tag{2}$$

such that

$$|H_k(j\omega)| = \frac{a_k\omega^4 + b_k\omega^2 + c_k}{\omega^4 + \alpha_k\omega^2 + \beta_k} \approx f_k(\omega) \tag{3}$$

where α_k and β_k are real and positive and $f_k(\omega)$ is a specified magnitude function which is assumed to be continuous in $0 < \omega < 1$.

The polynomial $N_k(s)N_k(-s)$ can be written as

$$N_k(s)N_k(-s)]_{j\omega} = [\omega^4 - 2(\eta_k^2 - \sigma_k^2)\omega^2 + (\eta_k^2 + \sigma_k^2)^2] \qquad (4)$$

where η_k and σ_k are shown in Fig. 2.

From equations (2) and (4) we obtain:

$$\frac{b_k}{a_k} = \xi_k = -2(\eta_k^2 - \sigma_k^2), \quad \frac{c_k}{a_k} = \tilde{\xi}_k^2 = (\eta_k^2 + \sigma_k^2)^2 \qquad (5)$$

From Equation (5) we see that $\frac{c_k}{a_k}$ must be positive, which implies that

$$a_k > 0 \qquad (5a)$$

since $c_k > 0$.

Again, from Equation (5) it may be shown that

$$\sigma_k = \pm\sqrt{\frac{1}{2}(\tilde{\xi}_k - \frac{\xi_k}{2})} \quad \text{and} \quad \eta_k = \pm\sqrt{\frac{1}{2}(\tilde{\xi}_k + \frac{\xi_k}{2})} \qquad (5b)$$

Now, since $\tilde{\xi}_k = \sigma_k^2 + \eta_k^2$ is always positive, we require $|\xi_k| < 2\tilde{\xi}_k$ in order that σ_k and η_k be real.

That is, $\left|\frac{b_k}{a_k}\right| < 2\sqrt{\frac{c_k}{a_k}}$. (6)

Equations (5a) and (6) are the conditions to be satisfied by $|H_k(j\omega)|$ in order that the corresponding $H_k(s)$ be realizable.

APPROXIMATION PROCEDURE

The approximation

$$\frac{a_k\omega^4 + b_k\omega^2 + c_k}{\omega^4 + \alpha_k\omega^2 + \beta_k} \approx f_k(\omega) \qquad (7)$$

yields $c_k = \beta_k f_k(0)$ where $f_k(0)$ is a specified zero frequency magnitude. Therefore c_k is a constant which depends on β_k and $f_k(0)$. This leads us to minimize the following error of approximation in the mean-square-error sense subject to the constraint that $c_k = C$ is a constant:

$$E \,\alpha \int_0^{w_0} [\frac{a_k\omega^4 + b_k\omega^2 + C}{\omega^4 + \alpha_k\omega^2 + \beta_k} - f_k(\omega)]^2 \qquad (8)$$

where E is the mean-square error. If we normalize $\omega_0 = 1$, then we get from Equation (8):

$$E \,\alpha \int_0^1 \frac{(a_k\omega^4 + b_k\omega^2 + C)^2}{(\omega^4 + \alpha_k\omega^2 + \beta_k)^2} d\omega - 2\int_0^1 \frac{(a_k\omega^4 + b_k\omega^2 + C)}{(\omega^4 + \alpha_k\omega^2 + \beta_k)} f_k(\omega)\, d\omega + \int_0^1 f_k^2(\omega)\, d\omega \qquad (9)$$

The integral $\int_0^1 f_k^2(\omega)d\omega$ is merely a constant.

We use the following notation:

$$I_2^{(n)} = \int_0^1 \frac{\omega^n}{(\omega^4 + a_k\omega^2 + \beta_k)^2} d\omega, \quad I_2^{(n,f_k)} = \int_0^1 \frac{\omega^n f_k(\omega)}{(\omega^4 + a_k\omega^2 + \beta_k)^2} d\omega$$

$$I_1^{(n)} = \int_0^1 \frac{\omega^n}{(\omega^4 + a_k\omega^2 + \beta_k)} d\omega, \quad I_1^{(n,f_k)} = \int_0^1 \frac{\omega^n f_k(\omega)}{(\omega^4 + a_k\omega^2 + \beta_k)} d\omega \qquad (10)$$

Then it follows from Equation (9) that

$$E \propto a_k^2 I_2^{(8)} + b_k^2 I_2^{(4)} + C^2 I_2^{(0)} + 2a_k b_k I_2^{(6)} + 2b_k C I_2^{(2)} +$$
$$+ 2a_k C I_2^{(4)} - 2[a_k I_1^{(4,f_k)} + b_k I_1^{(2,f_k)} + C I_1^{(0,f_k)}] + \int_0^1 f_k^2(\omega)d\omega \qquad (11)$$

The above integrals may be evaluated using integral tables.[2] Alternatively, Simpson's rule can be used to good advantage as indicated in the numerical example which illustrates this procedure.

To calculate a_k and b_k we set

$$\frac{\partial E}{\partial a_k} = \frac{\partial E}{\partial b_k} = 0$$

obtaining

$$a_k I_2^{(8)} + b_k I_2^{(6)} = I_1^{(4,f_k)} - C I_2^{(4)}$$

$$a_k I_2^{(6)} + b_k I_2^{(4)} = I_1^{(2,f_k)} - C I_2^{(2)} \qquad (12)$$

C is chosen such that the approximation is within the specifications of the magnitude characteristic at zero frequency and hence Equation (12) yields a_k and b_k. Assuming that a_k so obtained is positive, we adjust b_k (if necessary) such that the condition in Equation (6) is satisfied. If a_k turns out to be negative, repeat Equations (11), (12), and (13) with a different choice of H_k. See Table 1. The resulting $|H(j\omega)| = \frac{a_k\omega^4 + b_k\omega^2 + C}{\omega^4 + a_k\omega^2 + \beta_k}$ yields a realizable $H(s)$.

The above approximation may be refined by using sensitivity techniques which involve shifting the zeros of $H(s)$ in successive steps. This is equivalent to perturbing b_k for a fixed C. This procedure leads to an optimal set of values of b_k and C. These concepts are best illustrated by the following example:

Example. It is desired to obtain $H(s)$ which possesses a maximally flat time-delay and a magnitude characteristic as shown in Fig. 3 within approximately ± 1 db.

Here $f(\omega) = (1-\omega)$, $0 \le \omega \le 1$.

$D(s) = s^2 + 6s + 12$ is a second order Bessel polynomial which has a maximally flat time delay $= \frac{1}{2}$.

We seek $|H(j\omega)| = \dfrac{a\omega^4 + b\omega^2 + C}{\omega^4 + 12\omega^2 + 144}$

The integrals in Equation (12) are evaluated using Simpson's rule. For this example, two equally spaced intervals are found sufficient to yield satisfactory accuracy and thus we have$^{(3)}$

$$\int_0^1 g(\omega)d\omega \le \frac{h}{3}[g(\omega_0) + 4g(\omega_1) + g(\omega_2)] \tag{13}$$

where $h = 0.5$ with $\omega_0 = 0$, $\omega_1 = 0.5$ and $\omega_1 = 1$.

The corresponding error estimate is

$$\varepsilon = \frac{-h}{90}\frac{d^4 g(\xi)}{d\omega^4}, \quad 0 \le \xi \le 1 \tag{14}$$

The computations yield

$$\begin{array}{ll}
I_1^{(2)} = 0.2195 \times 10^{-2} & I_2^{(2)} = 0.14467 \times 10^{-4} \\
I_1^{(3)} = 0.1628 \times 10^{-2} & I_2^{(4)} = 0.08687 \times 10^{-4} \\
I_1^{(4)} = 0.1345 \times 10^{-2} & I_2^{(6)} = 0.07243 \times 10^{-4} \\
I_1^{(5)} = 0.12032 \times 10^{-2} & I_2^{(8)} = 0.06882 \times 10^{-4} \\
I_1^{(4,f)} = I_1^{(4)} - I_1^{(5)} & \\
I_1^{(2,f)} = I_1^{(2)} - I_1^{(3)} &
\end{array} \tag{15}$$

Substituting for $I_{1,2}^{(n)}$ in Equation (14), we get

$$\begin{aligned}
0.07243a + 0.08687b &= 5.6663 - 0.14467C \\
0.06882a + 0.07243b &= 1.417 - 0.08687C
\end{aligned} \tag{16}$$

Suppose we choose $C = 135$ which yields a zero frequency deviation within 1 db. Then Equation (16) yields:

$$a = 141.0287, \quad b = -276.338.$$

This gives $\sqrt{\frac{c}{a}} = \eta^2 + \sigma^2 = 0.98$.

And $\frac{1}{2}\left|\frac{b}{a}\right| = \eta^2 - \sigma^2 = 0.9797$, which suggests that $\eta^2 - \sigma^2$ be reduced slightly. This, in effect, increases σ. Thus, consider

$$\eta^2 + \sigma^2 = 0.98$$

$$\eta^2 - \sigma^2 = 0.96 \qquad (17)$$

from which

$$\eta = 0.986 \quad \text{and} \quad \sigma = 0.1$$

Correspondingly, $|b| = 2(0.96)a = 270.775$

Thus $$|H_1(j\omega)| = \frac{141.0287\omega^4 - 270.775\omega^2 + 135}{\omega^4 + 12\omega^2 + 144} \qquad (18)$$

i.e. $$H_1(s) = 141.0287 \frac{(s^2 + 0.1s + 0.98)(s^2 - 0.1s + 0.98)}{(s^2 + 6s + 12)^2}$$

The plot of $|H_1(j\omega)|$ is given in Fig. 4 and meets the required specifications.

Similarly $\eta^2 - \sigma^2 = 0.94$ would give $\sigma = 0.1415$ and

$$H_2(s) = \frac{141.0287(s^2 + 0.1415s + 0.98)(s^2 - 0.1415s + 0.98)}{(s^2 + 6s + 12)^2}$$

The plot of $|H_2(j\omega)|$ is shown in Fig. 4.

Fig. 4 shows the magnitude plots with $b = 135$ for different values of η. These plots clearly indicate that digital techniques may be used to good advantage to obtain the optimal realizable $H(s)$.

Generaliztions. A desired $H(s)$ may be written as a product

$$H(s) = \prod_{i=1}^{N} H_k(s)$$

where

$$H_k(s) = \frac{N_k(s)N_k(-s)}{D_k^2(s)}$$

and $T_k(s) = 1/D_k(s)$ is an all-pole function that has a maximally flat time delay. We refer to the H_k as "auxiliary functions." Thus, for each desired function f_k we seek H_k which approximates it.

Fig. 5(a) shows the overall desired magnitude function of f_1. Then $|H_1(j\omega)|$ is obtained to approximate f_1 by the method outlined above. Next, f_2 is chosen such that $f_1 = f_2 \cdot |H_1(j\omega)|$. This is done by sampling (Fig. 5(b)) and by conventional interpolation methods.[5] Thus, f_2 is obtained as a polynomial. Again, $|H_2(j\omega)|$ is obtained in the same manner as $|H_1(j\omega)|$. Then we have

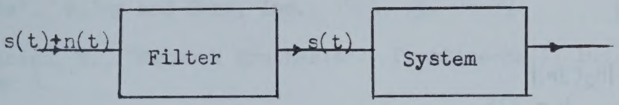
$$f_3 = \frac{f_1}{|H_1(j\omega)||H_2(j\omega)|}$$

(See Fig. 5(c)). This procedure is repeated until $H(s) = \prod_{i=1}^{N} H_k(s)$ approximates the given f_1 within ±n db. as per specifications. Note that each $H_k(s)$ is realizable.

We list the results that pertain to the initial approximation together with the remaining possible choices of $H_k(j\omega)$ in Table 1.

TABLE 1

$H_k(j\omega)$	Evaluation of coefficients	Realizability criteria
(a) $\dfrac{b_k\omega^2 + C}{\omega^2 + \gamma_k}$	$b_k = \dfrac{\tilde{I}_1^{(2,f)} - C\tilde{I}_2^{(2)}}{\tilde{I}_2^{(4)}}$	$b_k > 0$
	$\tilde{I}_1^{(2,f)} = \int_0^1 \dfrac{\omega^2 f(\omega)\,d\omega}{\omega^2 + \gamma_k}$	
	$\tilde{I}_2^{(2)} = \int_0^1 \dfrac{\omega^2\,d\omega}{(\omega^2 + \gamma_k)^2}$	
	$\tilde{I}_2^{(4)} = \int_0^1 \dfrac{\omega^4\,d\omega}{(\omega^2 + \gamma_k)^2}$	
(b) $\dfrac{b_k\omega^2 + C}{\omega^4 + a_k\omega^2 + \beta_k}$	$b_k = \dfrac{\hat{I}_1^{(2,f)} - C\hat{I}_2^{(2)}}{\hat{I}_2^{(4)}}$	$b_k > 0$
	$\hat{I}_1^{(2,f)} = \int_0^1 \dfrac{\omega^2 f(\omega)\,d\omega}{\omega^4 + a_k\omega^2 + \beta_k}$	
	$\hat{I}_2^{(2)} = \int_0^1 \dfrac{\omega^2\,d\omega}{(\omega^4 + a_k\omega^2 + \beta_k)^2}$	
	$\hat{I}_2^{(4)} = \int_0^1 \dfrac{\omega^4\,d\omega}{(\omega^4 + a_k\omega^2 + \beta_k)^2}$	
(c) $\dfrac{a_k\omega^4 + b_k\omega^2 + C}{\omega^4 + a_k\omega^2 + \beta_k}$	Solve: $a_k I_2^{(8)} + b_k I_2^{(6)} = I_1^{(4,f)} - CI_2^{(4)}$ $a_k I_2^{(6)} + b_k I_2^{(4)} = I_1^{(2,f)} - CI_2^{(2)}$ where the integrals are defined in Equation (10).	$a_k > 0$ $\left\lvert \dfrac{b_k}{a_k} \right\rvert < 2\sqrt{\dfrac{C}{a_k}}$


s(t)+n(t)
Filter
s(t)
System

s(t): signal. n(t): noise

Fig. 1(a)

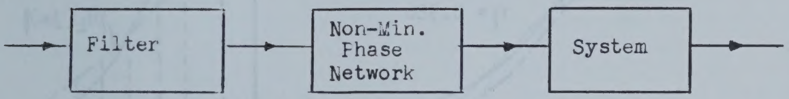

Filter
Non-Min.
Phase
Network
System

Fig. I(b)

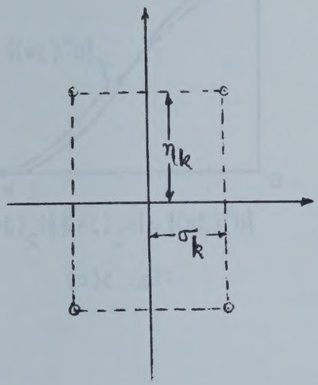

η_k
σ_k

Fig. 2

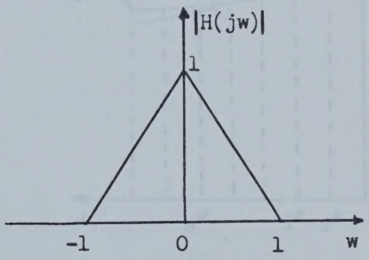


Fig. 3

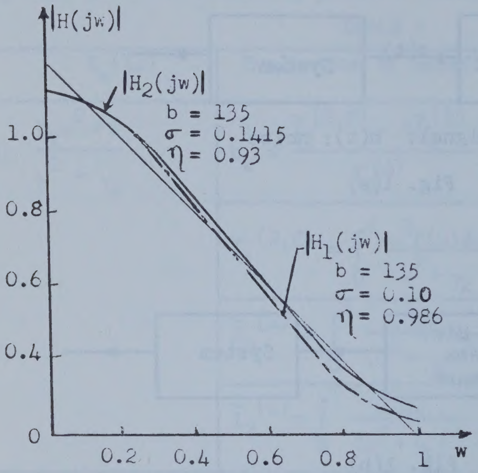


Fig. 4

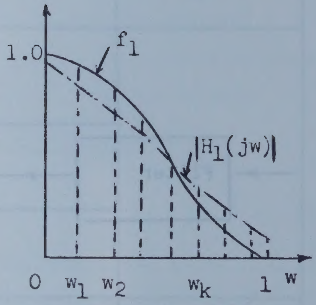


Fig. 5(a)

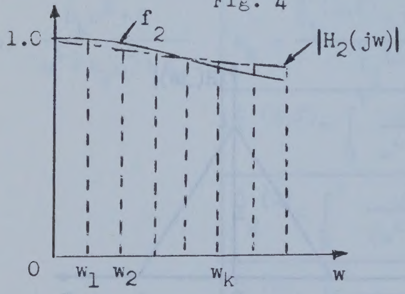
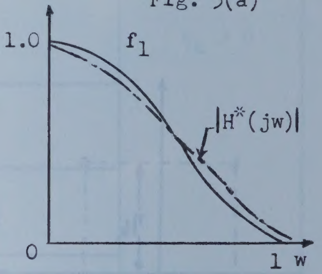


Fig. 5(b)



$|H^{*}(jw)| = |H_1(jw)|\,|H_2(jw)|$

Fig. 5(c)

REFERENCES

1. Calahan, D.A. "Modern Network Synthesis", Vol. 1: Approximation. Hayden Book Co., Inc., New York, 1964, pp. 11-12.

2. Bois, G.P., "Tables of Indefinite Integrals". Dover Publications Inc., New York, 1961, pp. 16-106.

3. Carnahan, B., Luther, H.A., Wilkes, J.O., "Applied Numerical Methods". Wiley and Sons, Inc., 1964, pp. 98-99.

4. Balabanian, N., "Network Synthesis". Prentice-Hall, Inc., 1958, Chapter 8.

5. Wilf, H.S., "Curve-Fitting Matrices", AMM 65, 1958, pp. 272-274.

IMPROVED ACTIVE REALIZATION OF ELLIPTIC-FUNCTION FILTERS

P. M. LIN
School of Electrical Engineering
Purdue University
Lafayette, Indiana

ABSTRACT

This paper presents some new approaches to active RC filter synthesis with the aim of reducing the number of elements. Improvements are brought about through

(1) the use of finite loss elliptic-function approximations, and

(2) the incorporation of a simplified passive RC transfer function synthesis technique, due to Lin and Siskind, with Linvill's active synthesis technique.

The main results of the paper may be summarized as follows: For low-pass filter specifications which normally require a 4th order elliptic-function approximation, the present method requires only one NIC, five capacitances, and eight resistances. Two of the eight resistances may be used as terminating resistances. The five capacitances needed in the realization are only one in excess of the theoretical minimum of four capacitances. More generally, for low-pass specifications which normally require a 4mth order elliptic-function approximation, the present method requires m cascaded NIC sections separated by isolating amplifiers. This is half the number of sections required by the conventional cascade synthesis which uses one NIC section for every pair of complex transmission zeros.

ULTRASPHERICAL FILTERS - TIME AND FREQUENCY RESPONSES

D. E. JOHNSON and J. R. JOHNSON
Dept. of Electrical Engineering
Louisiana State University
Baton Rouge, Louisiana

ABSTRACT

The problem of approximating the ideal normalized amplitude response of a low-pass filter by the use of a set of ultraspherical polynomials is considered. The amplitude response obtained is more general than the analogous response of the Chebyshev filter because of an additional parameter available with the ultraspherical polynomials. It is shown that this additional parameter may be used to obtain a response having either less ripple or sharper cut-off than the Chebyshev response. Comparisons are also made with the Butterworth filter, showing that for essentially the same results the ultraspherical filter can be constructed with considerably fewer elements. The phase, impulse, and step responses are also considered, and it is shown that the additional parameter may be utilized to improve the filters in these respects.

TIME DOMAIN SYNTHESIS OF A CLASS OF RLC NETWORKS

DONALD M. LEVY
Dept. of Electrical Engineering
University of Iowa
Iowa City, Iowa

DAVID P. BROWN
Dept. of Electrical Engineering
University of Wisconsin
Madison, Wisconsin

ABSTRACT

The time domain description of a class of RLC graphs is given in terms of first-derivative-explicit system of differential equations. In particular, the class of RLC graphs discussed have C-elements in a tree and R- and L-elements in the cotree of the graph, and further, have no all C-circuits and no all L-segs. Given one or two real square matrices, necessary and sufficient conditions for the synthesis of connected n+1 vertex RLC graphs with a star tree and a linear tree are derived, such that the matrices are coefficient matrices in the systems of differential equations characterizing the RLC graphs. The case where one coefficient matrix is specified is shown to be equivalent to synthesis based upon Bashkow's A-matrix for the RLC graph.

SUBOPTIMAL CONTROL BY FREQUENCY DOMAIN TECHNIQUES, DIFFERENTIAL APPROXIMATION, AND QUASILINEARIZATION

ANDREW P. SAGE
Department of Electrical Engineering
University of Florida
Gainesville, Florida

ABSTRACT

A theory of suboptimal linear control for systems with random or deterministic inputs is developed in this paper. The method consists in part of an approximation to the infinite time Kalman gains of the filtering or regulator problem. This is quickly accomplished on a Bode diagram by employing the condition that the return difference of the optimal system be greater than unity at all frequencies. The solution is then extended to the finite time regulator or nonstationary filtering problem posing specific suboptimal solutions to the Riccati equation and a cost function based on a quadratic measure of error between the true and the suboptimal solution. The methods of differential approximation and quasilinearization are used in conjunction with the frequency domain technique or by themselves in order to minimize the cost function and thereby obtain the suboptimal control. As a result, on line computer requirements to generate the time varying gains are kept to a minimum.

INTRODUCTION, FREQUENCY DOMAIN FILTERING AND CONTROL

Considerable progress has been achieved in recent times in the development of the analytical theory of linear control. Serious obstacles to the practical application of this theory exist because of the mathematical complexity necessary in applying the solution technique to problems involving a large number of state variables.

A theory of suboptimal control and filtering for processes with random or deterministic inputs is developed in this paper. The method consists in part of approximation to the infinite time Kalman gains for either the filtering or the regulator problem.

This is accomplished by noting that the minimization of the integral (2)

$$J_1 = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \{ \Phi(s) [1 - W(s)] [1 - W(-s)] + W(s)W(-s) \} \, ds \quad (1)$$

in which $\Phi(s)$ is a rational spectral density and $W(s)$ a realizable transfer function is accomplished by an open loop transfer function

$$G(s) = \frac{W(s)}{1-W(s)} = [\Phi(s) + 1]_+ - 1 \quad (2)$$

Where $[\ ]_+$ is the spectral factorization operator denoting the extraction of the multiplicative term containing poles and zeros in the left half plane only. For the case where

$$\Phi(s) = \frac{1}{(-1)^n (s)^{2n}} \quad (3)$$

the open loop transfer function is approximately

$$G(s) \approx \frac{B_{n-1} (\sqrt{2}\, s)}{s^n} \quad (4)$$

where $B_n(s)$ is the nth order normalized Butterworth polynomial. In general, it can be concluded that an approximation to the optimum G(s) is

$$G(s) \approx [\Phi(s)]_{+} \, B_{n-1}\left(\sqrt{2}\,\frac{s}{\omega_c}\right) \qquad (5)$$

Where ω_c is the frequency where $\Phi(s) = 1$. This construction can be conveniently and quickly accomplished on a Bode amplitude diagram or by the digital simulation of the above equations. Examination of the approximation errors involved in this expression have shown them to be quite small.

The optimal linear time invariant regulator problem consists of obtaining a continuous control law $\mu_o(t)$ for the system described by

$$\frac{dx}{dt} = Ax + b\,\mu(t) \qquad (6)$$

$$X(o) = Xo$$

$$z = Cx \qquad (7)$$

where b, x and z are n vectors, $\mu(t)$ is the control function, A and C are constant n x n matrices. The system is completely controllable and observable. The control law is chosen in such a fashion that the integral

$$J_2 = 1/2 \int_o^T \{ z'Qz + r\,\mu^2(t)\} \, dt + 1/2\, z'(T)\, Sz(T) \qquad (8)$$

is minimized. For the infinite time interval case this integral expression has exactly the same form as equation (1) and therefore the optimum solution is

$$k_\infty' \Phi(s)\, b = \{ 1 + \| \Phi(s)\, b \|^2_{\frac{C'QC}{r}} \}_{+}^{-1} \qquad (9)$$

$$\text{where } \mu_o(t) = k_\infty' x = k' \; C^{-1} z \qquad (10)$$

$$\Phi(s) = (sI-A)^{-1} \qquad (11)$$

The approximate frequency domain technique is capable of solving the infinite time regulator problem in a rapid and accurate manner, even when a large number of state variables are present. The same can be said of the white noise stationary filtering problem.

The "time domain" solution(3) to equations 6, 7, and 8 for the finite time of control problem is

$$\mu_o(t) = -\frac{bP(t)\, Cx}{r} = k(t)x = k(t)\, C^{-1} z \qquad (12)$$

where

$$\frac{-dP}{dt} = A'P + PA - \frac{Pgg'P}{r} + C'QC \qquad (13)$$

$$P(T) = S \qquad (14)$$

Equation 13 is the Matrix Riccati equation. In cases where terms of the S matrix are large, it is convenient to work with the inverse Riccati equation.

$$\left(\frac{dP}{dt} \right)^{-1} = P^{-1}A + A'P^{-1} - bR^{-1}b^{T} - P^{-1}c^{T}QC^{-1}P \qquad (15)$$

$$P^{-1}(T) = S^{-1} \qquad (16)$$

An alternate approach would be to rewrite the performance index as

$$J_2 = 1/2 \int_{o}^{T} \{ z'(t)'Qz(t) + r\mu^2(t) + \frac{d}{dt}[z'(t)Sz(t)] \} dt \qquad (17)$$

and derive a Riccati equation solution to this problem which would have the final condition $P(T) = o$ independent of S.

Since The White Measurement Noise Filtering Problem is the dual of the Regulator Problem, the same technique can be used for each problem. The optimal linear White Measurement Noise Filtering Problem consists of obtaining a linear least squares estimate of a state vector x based on observations z where

$$\dot{x} = Ax + Bw(t) \qquad (18)$$

$$z = cx + v(t) \qquad (19)$$

Where x and c are n vectors, w is an m vector, z, and v are scalars, A, B, are time varying matrices of compatible order. In addition, w(t) and v(t) are independent Gaussian white noise processes with zero means and covariance matrices

$$\text{Cov}\,[w(t), w(\gamma)] = Q\,\delta(t-\gamma) \qquad (20)$$

$$\text{Cov}\,[v(t), v(\gamma)] = r\,\delta(t-\gamma) \qquad (21)$$

$$\text{Cov}\,[w(t), v(\gamma)] = 0 \qquad (22)$$

for all t, γ, where δ is the Dirac delta function, Q is a symmetric $m \times m$ nonnegative definite matrix and r is a scalar. The measurement $z(t)$ starts at some time t_o and $E\{x(t_o)\} = 0$, where E denotes the expected value. Also, it is assumed that $\text{cov}\{x(t_o), x(t_o)\}$ is known.

The optimal estimate denoted by $\hat{x}$ of the state x at time $t > t_o$, such that $E\{\| x(t_1) - \hat{x}(t_1|t) \|\}$ is minimum and $E\{\hat{x}(t_1|t)\} = E\{x(t_1)\}$ given observations up to time t, is generated by the linear system (3)

$$\dot{\hat{x}}(t) = A\hat{x}(t) + K(t)[z(t) - c\hat{x}(t)] \qquad (23)$$

with

$$k(t) = \frac{P(t)\,c'}{r} \qquad (24)$$

and

$$\dot{P} = AP + PA' - Pc'r^{-1}cP + BQB' \qquad (25)$$

where

$$\hat{x}(t_o) = E\{x(t_o)\} = 0 \qquad (26)$$

$$P(t_o) = \mathrm{Cov}\,[x(t_o), x(t_o)] \qquad (27)$$

For the stationary case ($t_1 \gg t_o$), equations 18 and 19 can be written in terms of spectral densities as

$$\Phi_{xx}(s) = \Phi(s)\, BQB'\, \Phi'(-s) \qquad (28)$$

$$\Phi_{zz}(s) = c\, \Phi_{xx}(s)\, c' + r \qquad (29)$$

where $\Phi(s) = (sI - A)^{-1}$ and Φ_{xx} and Φ_{zz} are the Fourier transforms of the matrices Cov $[x(t), x(t+\gamma)]$ and Cov $[z(t), z(t+\gamma)]$.

The mean square error in estimation of $x(t)$, $E\{\| x - \hat{x} \|\}$ is minimized by the Kalman gains

$$k'_\infty\, \Phi(s)\, c' = \left\{1 + \frac{c\, \Phi_{xx}(s)\, c'}{r}\right\}_+^{-1} \qquad (30)$$

where

$$\hat{x}(s) = [sI - A + k_\infty c]^{-1} k_\infty z \qquad (31)$$

or its equivalent, equation 23, specifies the form of the optimum filter. In order to solve nonstationary problems the solution to equation 25 is approximated and then joined to the stationary (infinite time) Kalman gains through the use of appropriate variational methods.

ERRORS IN SUBOPTIMAL FILTERING

The error in the suboptimal estimate of $\hat{x}$, which will be denoted by $\tilde{y}$, is given by

$$\tilde{y} = x(t) - \hat{x}_{so}(t) \qquad (32)$$

Where $\hat{x}_{so}(t)$ is the suboptimal estimate of x. The derivative of this expression is

$$\dot{\tilde{y}} = \dot{\tilde{x}} - \dot{\hat{x}}_{so}(t) \qquad (33)$$

substituting from 18, 19, and 23

$$\dot{\tilde{y}} = Ax + Bw(t) - A\hat{x}_{so} + k_{so}(t)\,[cx + v(t) - c\,\hat{x}_{so}] \qquad (34)$$

By regrouping terms, equation 34 becomes

$$\dot{\tilde{y}} = [A - k_{so}(t)\, c]\; [x - \hat{x}_{so}] + k_{so}(t)\, v(t) + Bw(t) \qquad (35)$$

$$\dot{\tilde{y}} = D(t)\, \tilde{y} - k_{so}(t)\, v(t) + Bw(t) \qquad (36)$$

where

$$D(t) = A - k_{so}(t)\, c \qquad (37)$$

Now let $E(t)$ denote the covariance matrix of $\tilde{y}(t)$ which is given by

$$E(t) = \mathrm{Cov}\,[\tilde{y}(t), \tilde{y}'(t)] \qquad (38)$$

It is not difficult to show that the differential equation for the covariance of the error in the suboptimal estimate as a function of the original process dynamics and the suboptimal filter gains is

$$\dot{E} = D(t) E + ED^T (t)$$

$$+ k_{so}(t) \, r \, k'_{so}(t) + BQB' \qquad (39)$$

By use of equation 39 error analysis of the suboptimal (and optimal) filters can be conducted. Two approximations are involved here. One is the approximation of the infinite frequency Kalman gains. The other is the approximation to the solution to the Riccati equation which will now be discussed.

QUASILINEARIZATION AND DIFFERENTIAL APPROXIMATION

For problems where the number of state variables is large, both the dimension and the number of matrix operations involved in the real time computer synthesis of the Kalman gains becomes large. Variational methods may be used in order to approximate the Kalman gains by functions which can be generated either with less hardware or faster in real time problems. Two particularly appropriate methods are quasilinearization and differential approximation. The Kalman gains will be approximated by the gain Vector $k^*(t)$ which will be the solution to the n vector differential equation

$$\dot{k}^*(t) = f(k^*, \Gamma) \qquad (40)$$

$$\dot{\Gamma}(t) = 0 \qquad (41)$$

Where the form of the matrix differential equation describing k^* is known except for a certain number, m, of constant parameters of Γ, which are to be determined such that either

$$J_3 = 1/2 \int_0^T \| k(t) - k^*(t) \|^2_G \, dt \qquad (42)$$

or

$$J_4 = 1/2 \int_0^T \| \dot{k}(t) - f(k, \Gamma) \|^2 \, dt \qquad (43)$$

is minimum. Either the initial or the final value of $k^*(t)$ will be known. For the regulator problem

$$k^*(T) = -\frac{bSC}{r} \qquad (44)$$

whereas for the filter problem

$$k^*(t_o) = \frac{P(t_o)c}{r} \qquad (45)$$

For convenience, the development which follows will concentrate on the filter problem and thus the initial conditions of equation 45. The regulator problem is treated in an entirely analogous fashion.

Minimization of equation 42 subject to the constraints of equations 40 and 41 and the initial conditions of equation 45 can be accomplished by classical Euler Lagrange Methods or The Maximum Principle. The optimum parameter vector, Γ, is determined by solution of differential equations 40 and 41 with the initial conditions of equation 45 and the adjoint vector differential equations

$$\dot{\lambda}_k(t) = -[\nabla_k f(k^* \Gamma)]\lambda_k - G[k^*(t) - k(t)] \qquad (46)$$

$$\dot{\lambda}_\Gamma(t) = -[\nabla_\Gamma f(k^* \Gamma)]\lambda_k \qquad (47)$$

$$\lambda_\Gamma(T) = \lambda_\Gamma(t_o) = \lambda_k(T) = 0 \qquad (48)$$

Where ∇ represents the gradient operator. Equations 40, 41, 45, 46, and 47 represent a two point nonlinear boundary value problem which may be solved by quasilinearization (4, 5, 6). A new $2(n + m)$ vector

$$\eta'(t) = [k^*(t), \Gamma', \lambda'_k(t), \lambda'_\Gamma(t)] \qquad (49)$$

will be defined such that these equations may be described by the vector differential equation

$$\dot{\eta}(t) = g[\eta(t)] \qquad (50)$$

with boundary conditions

$$\langle d_i(jT), \eta(jT)\rangle = \alpha_i(jT) \qquad (51)$$

$j = 0, 1$

$i = 1, 2.....(m + n)$

where d and η are $2(m + n)$ dimensional vectors and $\langle\ ,\ \rangle$ denotes the inner product. If $\eta^o(t)$ is the initial guess to the solution of equation 50, the $(q + 1)$th approximation is obtained from the qth by

$$\dot{\eta}^{q+1}(t) = g[\eta^q(t)] + \mathcal{J}\{g[\eta^q(t)]\}\{\eta^{q+1}(t) - \eta^q(t)\} \qquad (52)$$

where $\mathcal{J}$ is the Jacobian Matrix whose ij^{th} element is the partial derivative $\partial g_i/\partial\eta_j$. If $\Phi^{q+1}(t)$ is the fundamental matrix of

$$\dot{\Phi}^{q+1}(t) = \mathcal{J}\{g[\eta^q(t)]\}\Phi^{q+1}(t) \qquad (53)$$

$$\Phi^{q+1}(o) = I = \text{identity matrix}$$

and $\Omega^{q+1}(t)$ is the particular vector solution of

$$\dot{\Omega}^{q+1}(t) = g[\eta^q(t)] + \mathcal{J}\{g[\eta^q(t)]\}\{\Omega^{q+1}(t) - \eta^q(t)\} \qquad (54)$$

$$\Omega^{q+1}(o) = o$$

The solution of equation 52 is

$$\eta^{q+1}(t) = \Phi^{q+1}(t)\,\nu^{q+1} + \Omega^{q+1}(t) \qquad (55)$$

where ν^{q+1} is a constant vector which is determined by solving

$$\langle d_i\,(jT),\ \Phi^{q+1}(jT)\ \nu^{q+1} + \Omega^{q+1}(jT)\rangle = \alpha_i(jT) \qquad (56)$$

$$j = 0,\ 1$$

$$i = 1,2,3,\ldots(m+n)$$

The quasilinearization method is readily programmed on a digital or hybrid computer and computation time is short as rapid convergence is assured by a good initial guess to $\eta^{o}(t)$ which is available from the solution to the Riccati equation in this case. An alternate approach consists in solving a quasilinearized version of equations 40, 41, and 45 and accomplishing a least squares curve fit to minimize equation 42 (7). This results in solving a multipoint (m+n) dimensional boundary value problem whereas the method proposed here results in a 2(m+n) dimensional 2 point boundary value problem. Both methods lead to equivalent results.

The optimum value of the parameter vector 17 in order to minimize equation 43 is the one which makes the partial derivatives of the integral with respect to the components of Γ equal to zero. This is

$$\int_o^T \mathcal{J}_\Gamma'\{\ f(k,\Gamma)\}\ \dot{k}\ dt = \int_o^T \mathcal{J}_\Gamma'\{\ f(k,\Gamma)\}\ \ f(k,\Gamma)\ dt \qquad (57)$$

where $\mathcal{J}_\Gamma'$ is the transpose of the Jacobian Matrix whose ijth element is the partial derivative $\partial fi/\partial \Gamma j$. Thus $\mathcal{J}_\Gamma'$ is a m by n matrix and solution of the m simultaneous equations of equation 52 yields the parameter vector . Computationally, this method of identification or approximation of system parameters is simpler to program and faster than the previous method. It has been given the name differential approximation,(8) and is an excellent starter for quasilinearization.

EXAMPLE OF NON-STATIONARY OPTIMIZATION

Consider a dynamic system in which the acceleration is white noise, a problem typical of aerospace guidance. Position is observed after it is contaminated by white noise. The process equation is therefore

$$\dot{x} = \begin{bmatrix} o & 1 \\ o & o \end{bmatrix} x + \begin{bmatrix} o \\ 1 \end{bmatrix} w(t)$$

where $x' = [\ x_1\ x_2\ ]$ and w(t) is a zero mean scalar Gaussian white noise with a constant variance which shall be taken equal to 1 for this example. The measurement is given by

$$z(t) = [\ 1 \quad o\ ]\ x(t) + v(t)$$

where z is a scalar and v(t) is a zero mean scalar Gaussian white noise with a constant variance of 16. It is further assumed that

$$P(o) = \begin{bmatrix} 1 & o \\ o & o \end{bmatrix}$$

The suboptimal solution of this filtering problem consists first of finding the stationary solution. Equation 28 becomes

$$\Phi_{xx}(s) = \begin{bmatrix} \frac{1}{s^4} & \frac{-1}{s^3} \\ \frac{1}{s^3} & \frac{1}{s^2} \end{bmatrix}$$

thus

$$\frac{c\,\Phi_{xx}(s)\,c'}{r} = \frac{1}{16\,s^4}$$

and

$$K'_{\infty}\,\Phi'(s)\,c' = \frac{1}{4\,s^2}\,[\,1 + 2\sqrt{2}s\,]$$

$$K'_{\infty} = [K_1\ K_2] = [\frac{1}{\sqrt{2}}\quad \frac{1}{4}\,]$$

This represents the suboptimal Kalman gain vector for "large" time. In this particular case, it is also the optimal large time gain vector. The optimal gain vector is mechanized as shown in Figure 1.

It can be shown that the eigenvalues of P(t) are composed entirely of the positive and negative zeros of

$$1 + \frac{c\,\Phi_{xx}(s)\,c'}{r} = 0$$

Thus the dominant time constants of P(t) would seem to be the inverse of the zeros of this equation. The exact solution to the nonstationary problem is obtained from the Riccati equation for this example.

$$\dot{P}_{11} = 2P_{12} - \frac{P_{11}^{\,2}}{16}$$

$$\dot{P}_{12} = P_{22} - \frac{P_{11}\,P_{12}}{16}$$

$$\dot{P}_{22} = \frac{-P_{12}^{\,2}}{16} + 1$$

and

$$K(t) = \begin{bmatrix} k_1(t) \\ \\ k_2(t) \end{bmatrix} = \frac{1}{16} \begin{bmatrix} P_{11}(t) \\ \\ P_{12}(t) \end{bmatrix}$$

The quasilinearization approximation will be obtained such that the following expression is minimum.

$$J = \frac{1}{2} \int_o^{10} \| K(t) - K^*(t) \|^2 \, dt$$

The suboptimal gains will satisfy the differential equation

$$\dot{k}^* = A [k^*(t) - k_\infty]$$

$$\dot{A} = o$$

where A is a 2 x 2 matrix of constant coefficients to be determined. For this particular example, the specific equations to be solved are

$$\dot{k}_1^* = a_{11} (k_1^* - \frac{1}{\sqrt{2}}) + a_{12} (k_2^* - \frac{1}{4})$$

$$\dot{k}_2^* = a_{21} (k_1^* - \frac{1}{\sqrt{2}}) + a_{22} (k_2^* - \frac{1}{4})$$

$$\dot{a}_{11} = \dot{a}_{12} = \dot{a}_{21} = \dot{a}_{22} = 0$$

$$-\dot{\lambda}_1 = a_{11}\lambda_1 + a_{21}\lambda_2 + k_1^* - k_1$$

$$-\dot{\lambda}_2 = a_{12}\lambda_1 + a_{22}\lambda_2 + k_2^* - k_2$$

$$-\dot{\lambda}_3 = (k_1^* - \frac{1}{\sqrt{2}}) \lambda_1$$

$$-\dot{\lambda}_4 = (k_2^* - \frac{1}{4}) \lambda_1$$

$$-\dot{\lambda}_5 = (k_1^* - \frac{1}{\sqrt{2}}) \lambda_2$$

$$-\dot{\lambda}_6 = (k_2^* - \frac{1}{4}) \lambda_2$$

with the boundary conditions

$$k_1^*(o) = \frac{1}{16}$$

$$k_2^*(o) = \lambda_1(10) = \lambda_2(10) = \lambda_3(o) = \lambda_3(10) = \lambda_4(o) = \lambda_4(10) = \lambda_5(o) = \lambda_5(10) = \lambda_6(o) = \lambda_6(10) = 0$$

This nonlinear two point boundary value problem is solved via the quasilinearization method. The first of the 12 coupled differential equations when quasilinearized yields

$$\dot{k}_1^{*q+1} = a_{11}^{q} \left(k_1^{*q} - \frac{1}{\sqrt{2}}\right) + a_{11}^{q} \left(k_1^{*q+1} - k_1^{*q}\right) + k_1^{*q} \left(a_{11}^{q+1} - a_{11}^{q}\right)$$

$$+ \; a_{12}^{q} \left(k_2^{*q} - \frac{1}{4}\right) + a_{12}^{q} \left(k_2^{*q+1} - k_2^{*q}\right) + k_2^{*q} \left(a_{12}^{q+1} - a_{12}^{q}\right)$$

Optimum values of the A matrix are

$$A = \begin{bmatrix} -1.95 & 5.57 \\ -.408 & .931 \end{bmatrix} \qquad A = \begin{bmatrix} .123 & 0 \\ 0 & -.423 \end{bmatrix}$$ (a_{12} and a_{21} are forced to zero)

The suboptimal and optimal Kalman gains computed from this method are shown in Figure 2. In order to obtain the suboptimal gain vector, it is first necessary to solve the matrix Riccati equation and to then solve a set of coupled 12th order differential equations in order to determine the parameter vector Γ (the A matrix in this example). Thus a healthy amount of off line computer solution effort is required in order to establish simpler on line requirements. The amount of off line computer solution time can be reduced considerably if the suboptimal gain differential equations are uncoupled.(9) This was the case in the second part of the example. Allowing the uncoupled suboptimal gain differential equations to become nonlinear or time varying would generally lead to a much better solution than the one presented here for the uncoupled case.

The differential approximation solution will be obtained such that the following expression is minimum

$$J = \frac{1}{2} \int_0^{10} \left\| \dot{k} - A\left[k(t) - k_\infty \right] \right\|^2 dt$$

where A is a 2 x 2 matrix of constant coefficients. For this specific example the optimum parameters are determined by simultaneous solution of the following linear algebraic equations.

$$a_{11} \int_0^{10} \left(k_1 - \frac{1}{\sqrt{2}}\right)^2 dt + a_{12} \int_0^{10} \left(k_2 - \frac{1}{4}\right)\left(k_1 - \frac{1}{\sqrt{2}}\right) dt = \int_0^{10} \dot{k}_1 \left(k_1 - \frac{1}{\sqrt{2}}\right) dt$$

$$a_{11} \int_0^{10} \left(k_1 - \frac{1}{\sqrt{2}}\right)\left(k_2 - \frac{1}{4}\right) dt + a_{12} \int_0^{10} \left(k_2 - \frac{1}{4}\right)^2 dt = \int_0^{10} \dot{k}_1 \left(k_2 - \frac{1}{4}\right) dt$$

$$a_{21}\int_0^{10}\left(k_1-\frac{1}{\sqrt{2}}\right)^2 dt + a_{22}\int_0^{10}\left(k_2-\frac{1}{4}\right)\left(k_1-\frac{1}{\sqrt{2}}\right) dt = \int_0^{10}\dot{k}_2\left(k_1-\frac{1}{\sqrt{2}}\right) dt$$

$$a_{21}\int_0^{10}\left(k_1-\frac{1}{\sqrt{2}}\right)\left(k_2-\frac{1}{4}\right) dt + a_{22}\int_0^{10}\left(k_2-\frac{1}{4}\right)^2 dt = \int_0^{10}\dot{k}_2\left(k_2-\frac{1}{4}\right) dt$$

The first two of the above equations are uncoupled from the last two. Optimum values of the A matrix are

$$A = \begin{bmatrix} -1.35 & 3.57 \\ -0.41 & 0.93 \end{bmatrix} \qquad A = \begin{bmatrix} -.213 & 0 \\ 0 & -.302 \end{bmatrix} \text{(a_{12} and a_{21} are forced to zero)}$$

Suboptimal and optimal Kalman gains computed from the differential approximation method are shown in Figure 3.

ACKNOWLEDGMENT

The author wishes to thank B. Eisenberg and S. Smith for assistance in the numerical experimentation phases of this work.

REFERENCES

1. Kalman, R.E., "When is a Linear Control System Optimal," Journal of Basic Engineering, American Society of Mechanical Engineers, March 1964.
2. Sage, A.P. and Streets, R.B., "Synthesis of Approximately Optimal Control Systems Using Bode Diagrams," Proceedings 1964 Allerton Conference on Circuit and System Theory, University of Illinois, October, 1964.
3. Kalman, R.E., T.S. Englar, and R.S. Bucy, "Fundamental Study of Adaptive Control Systems." ASD-TR 61-27, Wright Patterson Air Force Base, Ohio.
4. Bellman, R., H. Kagiwada, and R. Kalaba, "Quasilinearization, System Identification, and Prediction", RAND Memorandum RM-3812-PR, RAND Corp., August, 1963.
5. Prasanna Kumar, K.S., and R. Sridhar, "On the Identification of Control Systems by the Quasilinearization Method," IEEE Transactions on Automatic Control, Vol. AC-9, April, 1964.
6. Kalaba, R., "On Nonlinear Differential Equations, the Maximum Operation and Monotone Convergence", J. Math. Mech., Vol. 8, July, 1959.
7. Lavi, A., J.C. Strauss, "Parameter Identification in Continuous Dynamic Systems", 1965, IEEE International Convention Record, part 6, March, 1965.
8. Bellman, R., R. Kalaba and R. Sridhar, "Adaptive Control Via Quasilinearization and Differential Approximation", RAND Memorandum, RM 3928 PR, RAND Corp., November, 1963.
9. Meditch, J.S., "A Theory of Suboptimal Data Filtering for Space Navigation", SSD-TR-65-60, Space Systems Division, Air Force Systems Command, Los Angeles, California.

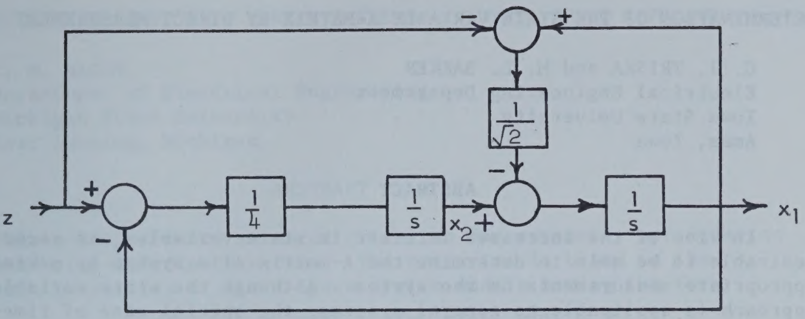


Figure 1. Block Diagram of Stationary Optimal Filter

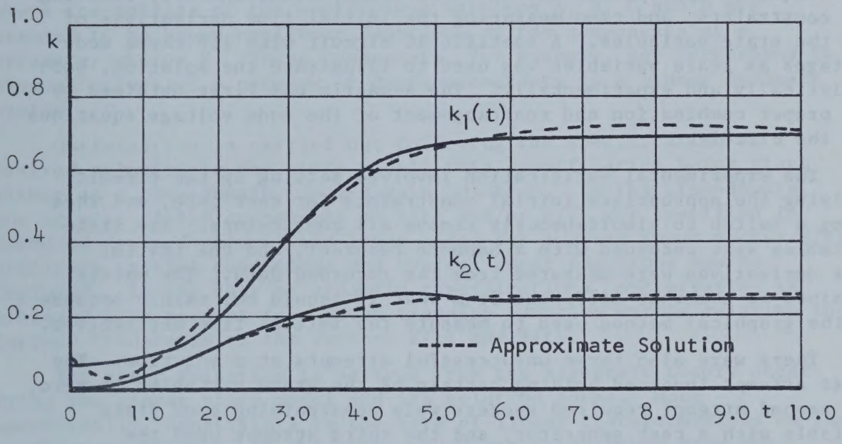


Figure 2. Results of Quasilinearization Approximation

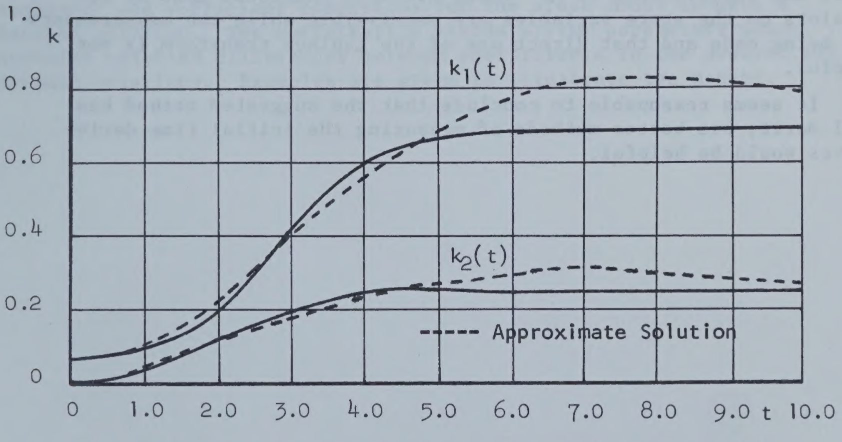


Figure 3. Results of Differential Approximation

DETERMINATION OF THE STATE VARIABLE A-MATRIX BY DIRECT MEASUREMENT

C. J. TRISKA and M. L. BAKKER
Electrical Engineering Department
Iowa State University
Ames, Iowa

ABSTRACT

In view of the increased interest in state variables, it seems desirable to be able to determine the A-matrix of a system by making appropriate measurements on the system. Although the state variable approach is applicable to general systems, the special case of time-invariant linear differential systems is considered.

Determination of the state variable A-matrix by direct measurement requires appropriate initial conditions, simultaneously removing all constraints, and then measuring the initial time derivatives of all the state variables. A specific RC circuit with its three node voltages as state variables was used to illustrate the solution, both analytically and experimentally. The A-matrix was first obtained by the proper combination and rearrangement of the node voltage equations for the circuit.

The experimental verification involved setting up the circuit, applying the appropriate initial constraints for each case, and then using a switch to simultaneously remove all constraints. The state variables were recorded with a Sanborn recorder, and the initial time derivatives were measured from the recorded data. The matrix obtained is somewhat different from what it should be, mainly because of the graphical method used to measure the initial time derivatives.

There were also three unsuccessful attempts at a solution. The first attempt involved holding certain of the state variables at zero, the second attempt required successively constraining each state variable with a test generator, and the third attempt used the Laplace transform to obtain the relationships between the initial state and the elements of A. These attempts indicated that no constraints on the state variables are permissible while the measurements are being made and that direct use of the Laplace transform is not helpful.

It seems reasonable to conclude that the suggested method has real merit, but better methods of measuring the initial time derivatives would be helpful.

PARAMETER OPTIMIZATION IN LINEAR SYSTEMS - A STATE-SPACE APPROACH

C. M. BACON
Department of Electrical Engineering
Michigan State University
East Lansing, Michigan

ABSTRACT

This paper presents an error criterion for parameter optimization of linear time-invariant systems. The criterion is based on the algebraic properties of state-space system models having the form

$$X(t) = A(P)\ X(t) + B(P)\ E(t)\,,\ X(0) = X_o$$

$$R(t) = C(P)\ X(t) + D(P)\ E(t)$$

where the entries of the coefficient matrices A, B, C and D are assumed to be known nonlinear functions of the components of a parameter vector P. These components, for example, may correspond to element values of a network or design constants in a linear feedback system.

Optimization is carried out following the specification of a desired solution to the state model; this specification being given either as a time-domain, or s-domain solution. In the time-domain, one or more of the components of X(t) and/or R(t) are specified in terms of desired natural and input modes for a given input vector E(t). Arbitrary initial conditions and typical input functions can be employed. In the case of s-domain specifications, optimization is achieved with respect to a given transfer function matrix relating the Laplace transforms of the vectors E(t) and R(t).

The error criterion is derived from the algebraic theory underlying the linear state model and its solution forms. More specifically, constraints in the form of linear transformations on the coefficient matrices of the state model are established from the solution specifications. These constraints have been shown to provide necessary and sufficient conditions for the state model to have a desired solution. The optimization yields system parameters which minimize weighted differences between coefficients in the desired and optimum solutions. Examples are given to illustrate the method.

IDENTIFICATION OF SYSTEM PARAMETERS BY IMPLICIT SYNTHESIS*

S. E. STEPHENSON
Department of Electrical Engineering
University of Arkansas
Fayetteville, Arkansas

ABSTRACT

A new technique is developed for the identification of the unknown parameters of a system which can be represented by a linear differential equation. The development of the technique is constrained by practical engineering criteria of equipment cost and size, in-service system parameter identification, and a potential for automation of the identification technique. The identification is based on an implicit synthesis technique in which an identifier circuit voltage is forced to a correct value by the circuit configuration. The identifier is composed of analog computing elements and operates on the input and output signals of the operating system.

The implicit synthesis technique is applied to identify two parameters simultaneously in a system described by a second-order differential equation. The identifier is tested on simulated dynamic systems to establish a parameter volume of identification and to investigate limitations imposed on the technique by input signal requirements. Successful identification is made for large variation in the system parameters. A brief investigation indicates that the technique may be extended to determine parameters of higher-order systems and to determine the unknown order of a system.

---

*The work in this paper was submitted in a dissertation to the Graduate College of the Texas A & M University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

THE CATHODE RAY STORAGE TUBE IN ADAPTIVE DECISION SYSTEMS

BAXTER F. WOMACK
The University of Texas
Austin, Texas

JAMES H. SEAMON
Missouri Pacific Railroad Company
Saint Louis, Missouri

ABSTRACT

An adaptive decision system is proposed using the cathode ray storage tube to realize as many as 10^6 adaptive weights. This type of system is sometimes called a learning machine or trainable pattern classification system. The hardware required for the proposed system consists of six interconnected units.

The logical structure of the system is two layer, with the first layer consisting of adaptive threshold logic units (TLUs) in parallel, and the second layer consisting of majority-rule TLUs in parallel. The interconnections of the layers are such that the system operates as parallel committees, and the committee selection procedure is used to select which TLUs are to be trained. The logic used is (1,0) which offers significant advantages in adapting the weights stored in the storage tube. The fixed increment training rule is used, and the size of the increment is selected by the operator.

The effects of limits on the dynamic range of the weights and of changes in increment size during training are investigated by simulations on a digital computer. In order to reduce the complexity and cost of the proposed system, a modified selection procedure for training is developed to replace the committee selection procedure. Both the committee procedure and modified procedure are used to train the simulated system.

Results indicate the hardware advantage for the modified procedure is offset by a shorter convergence time for the committee procedure. Results also indicate that restrictions on the dynamic range of the weights generally hinders the training operation, but that in some cases moderate restrictions on the dynamic ranges can help the training operation. Other results indicate that the initial size of the increment definitely affects the training operation. Further, the results indicate that the training operation is affected by changing the interaction at which the increment size is changed.

This paper will describe the proposed system and present the simulation results.

PLANAR EQUIVALENT NETWORKS FOR NON-PLANAR NETWORKS DERIVED BY MATRIX METHODS

ERWIN C. HAUSMANN and GLADWYN LAGO
Dept. of Electrical Engineering
University of Missouri
Columbia, Missouri

DOUGLAS H. DOBBS
Formerly Univ. of Missouri
Now U.S. Air Force

ABSTRACT

The concept of duality, as applied to electrical networks, has been the object of considerable study in recent years. It has been established that dual networks do not exist for any non-planar network. However, if a planar equivalent network can be found which is equivalent to the non-planar network as far as the characteristics at the external nodes are concerned, then the dual of this planar equivalent network can be considered to be an equivalent dual of the original non-planar network as far as the external characteristics are concerned. This paper is concerned with establishing a set of mathematical relationships between the elements of a non-planar network with n nodes, and the equivalent planar network of n+1 nodes, if such an equivalent planar network exists. Equivalent planar graphs are obtained for the four node complete graph and the five node complete graph. Equations relating the element values of the graphs are obtained by equating the nondiagonal entries of appropriate admittance matrices. Realizability of the elements of the equivalent planar graph is not considered. It is shown that a seven node equivalent planar graph can not be obtained for the six node regular graph of order three, commonly known as Kuratowski's six node graph, under the condition that the elements of the equivalent graph are constrained to remain within the perimeter of the original six nodes. A special case is solved which allows one element to exist outside the perimeter of the original six nodes. This special case can be used if two adjacent nodes, of the original six nodes, are not connected to any other part of a larger graph. Since it is shown that in the general case a seven node planar equivalent graph does not exist for this non-planar graph, the question of whether a planar equivalent graph of any number of nodes exists for this non-planar graph is considered next. A graphical analysis is carried out for several special cases by repeated application of partial subgraph replacement. The resulting equivalent graphs are all non-planar.

AN EQUIVALENT CIRCUIT FOR THE OPERATIONAL AMPLIFIER

EDWARD M. BUTLER and SANJIT K. MITRA
Cornell University
Ithaca, New York

ABSTRACT

In this paper, an equivalent circuit for the operational amplifier is proposed subject to the provision that the amplifier configuration in the pertinent circuit employs negative feedback. This condition is not restrictive, since virtually all circuits using such high gain devices must employ negative feedback for the purpose of stability. The equivalent circuit makes use of two so-called "pathological elements", the "nullator" and the "norator". The proposed equivalent representation allows the taking of duals of circuits containing operational amplifiers. This idea has led to a systematic approach to the realization of the four types of controlled sources. Another advantage of using the nullator-norator representation is in the analysis of operational amplifier circuits.

A CLASSIFICATION OF PR BIQUADRATIC IMMITTANCE FUNCTIONS AND THEIR REALIZATION USING A PERFECTLY COUPLED AND CENTER BRANCH AUGMENTED TWO-PORT

KURT H. HAASE
Air Force Cambridge Research Laboratories, L. G. Hanscom Field,
Bedford, Mass.

ABSTRACT

A positive real (pr) biquadratic immittance function is of the normalized form

$$F(s) = (s^2 + N_1 s + N_0)/(s^2 + D_1 s + D_0), \tag{1}$$

and its inverse is

$$F'(s) = 1/F(s) = (s^2 + N_1' s + N_0')/(s^2 + D_1' s + D_0'), \tag{2}$$

where

$$N_0' = D_0, \; N_1' = D_1, \; D_0' = N_0, \; D'_1 = N_1.$$

All function coefficients N_0, N_1, D_0, D_1 are positive and real. We refer to F(s) as a non-singular pr function if

$$N_1 D_1 > (\sqrt{N_0} - \sqrt{D_0})^2. \tag{3}$$

We refer to the function

$$\bar{F}(s) = (s^2 + \bar{N}_1 s + \bar{N}_0)/(s^2 + \bar{D}_1 s + \bar{D}_0) \tag{4}$$

as a singular pr function if

$$\bar{N}_1 \bar{D}_1 = (\sqrt{\bar{N}_0} - \sqrt{\bar{D}_0})^2. \tag{5}$$

A function F(s), Eq. (1), belongs to one of eight categories. The categories (a1), ..., (a4) are of the main category (a) for which $N_0/D_0 > 1$; the categories (b1), ..., (b4) are of the main category (b) for which $N_0/D_0 < 1$. Thus each main category is divided into four sub-categories (1), ..., (4). The category to which a particular function F(s), likewise F'(s), $\bar{F}(s)$ and $\bar{F}'(s)$, belongs can be recognized by investigating the sign polarities of three test values T_0, T_a, and T_1.

$$T_0 = N_0 - D_0, \tag{6}$$

$$T_a = D_0 T_0 - D_1(N_0 D_1 - N_1 D_0), \tag{7}$$

$$T_1 = T_0 - D_1(N_1 - D_1). \tag{8}$$

Eqs. (6,7,8) refer to F(s); for F'(s) and $\bar{F}(s)$ add prime and bar

notations. The following Table 1 presents all possible combinations of sign polarities and thus it gives the combination characterizing a certain category.

Test Value	Category (a1)	(a2)	(a3)	(a4)	(b1)	(b2)	(b3)	(b4)
T_0	+	+	+	+	-	-	-	-
T_a	-	+	+	-	+	-	+	-
T_1	-	+	-	+	+	-	-	+

Table 1

A function F(s) belongs to one main category, its inverse, the function F'(s) belongs to the other. Only if F(s) is of (a3), then F'(s) is of (b4), or if F(s) is of (b3), then F'(s) is of (a4). For all other categories the relations between the categories of F(s) and F'(s) are ambiguous as it is shown by Table 2.

A plot of the real component R_e F(jω) vs $\Omega = \omega^2$ is shown in Fig. 1 for the four sub-categories of the main category (a). A very similar picture exists for the representation of the real component for the main category (b). The ordinate N_0/D_0

If F(s) is of category	then F'(s) = 1/F(s) is of either on of the categories			
(a1)	(b1),			(b4)
(a2)		(b2),		(b4)
(a3)				(b4)
(a4)	(b1),	(b2),	(b3),	(b4)
(b1)	(a1),			(a4)
(b2)		(a2)		(a4)
(b3)				(a4)
(b4)	(a1),	(a2),	(a3),	(a4)

Table 2

is crossed at $\Omega_a = T_a/T_0$ in the categories (a2), (a3), (b2), and (b4). The ordinate 1 is crossed at $\Omega_1 = D_0T_0/T_1$ in the categories (a2), (a4), (b2), and (b3). The minimum in the magnitude of R_e F(jω)

is ≥ 1 at $\Omega = \infty$ in the categories (a1) and (a3); it is N_0/D_0 at $\Omega=0$ in the categories (b1) and (b3); it is $N_m < 1$ at Ω_m in the categories (a2), (a4), (b2), and (b4). Thus a function F(s) can be decomposed in a first step

if F(s) is of category (a1) or (a3)

$$F(s) = 1 + F_{ak}(s), \tag{9}$$

where

$$F_{ak}(s) = (N_1 - D_1)\ (s + b_1^2)/(s^2 + D_1 s + D_0), \tag{10}$$

with

$$b_1^2 = T_0/(N_1 - D_1); \tag{11}$$

if F(s) is of category (b1) or (b3)

$$F(s) = N_0/D_0 + F_{bk}(s), \tag{12}$$

where

$$F_{bk}(s) = (-T_0)s(s + b_1^2)/D_0(s^2 + D_1 s + D_0), \tag{13}$$

with

$$b_1^2 = (N_0 D_1 - N_1 D_0)/T_0. \tag{14}$$

When F(s) and F'(s) are of the categories (a1) and (b1) or vice versa, then $F_{ak}(s)$ and $F_{bk}(s)$ are immittance functions of two-kind element (RC or RL) ladders. The ladders for the impedance interpretation of F(s) = Z(s) are shown in the first and third row of Fig. 2. All circuits in a row are equivalences. If only one of the inverse functions is of the sub-category (1), or if they are of the categories (a3) and (b4) or (b3) and (a4) then $F_{ak}(s)$ and $F_{bk}(s)$ respectively are immittance functions of three-kind element (RLC) ladders. The ladders for the impedance interpretation of F(s) are shown in the rows 2 and 4 of Fig. 2. There are two equivalences in each row. The duals of the circuits in Fig. 2 hold for the admittance interpretation of F(s).

When F(s) is of one of the subcategories (2) or (4), then we decompose

$$F(s) = N_m + (1-N_m)\overline{F}(s) = N_m + \frac{1}{(1-N_m)'\,\overline{F}'(s)}, \tag{15}$$

where $\overline{F}(s)$ is of the same category as F(s) and $\overline{F}'(s) = 1/\overline{F}(s)$. In addition $(1-N_m)' = 1/(1-N_m)$. $\overline{F}(s)$ and $\overline{F}'(s)$ are singular. Eq. (15)

offers two equivalences depending on whether we use the primed or the unprimed values. For $\overline{F}(s)$ or $\overline{F}'(s)$ we set

$$\overline{F}(s) = \frac{vx(n-1)^2\varphi(s)\phi(s) + vzn^2\varphi(s)\xi(s) + xzn\phi(s)\xi(s)}{v\varphi(s) + xn\phi(s) + zn\xi(s)} \qquad (16)$$

where we identify $\varphi(s)$, $\phi(s)$, and $\xi(s)$ with any of the frequency functions s, 1/s, and 1. This allows six variations:

$\varphi(s) = s, \ \xi(s) = 1/s, \ \phi(s) = 1$ for $\overline{F}_a$,

$\varphi(s) = s, \ \phi(s) = 1/s, \ \xi(s) = 1$ for $\overline{F}_{a*}$,

---

$\varphi(s) = 1/s, \ \xi(s) = s, \ \phi(s) = 1$ for $\overline{F}_b$,

$\varphi(s) = 1/s, \ \phi(s) = s, \ \xi(s) = 1$ for $\overline{F}_{b*}$,

---

$\varphi(s) = 1, \ \phi(s) = s, \ \xi(s) = 1/s$ for $\overline{F}_c$,

$\varphi(s) = 1, \ \xi(s) = s, \ \phi(s) = 1/s$ for $\overline{F}_{c*}$.

Then we identify the constants v, n, x, z with the coefficients of $\overline{F}(s)$.

With

$U = u\varphi(s) = v(n-1)\varphi(s)$, $W = w\varphi(s) = v(\frac{1}{n} - 1)\varphi(s)$, $V = v\varphi(s)$, $X = x\phi(s)$, and $Z = z\xi(s)$ as impedances the circuit left in Fig. 3 has the driving-point impedance $Z(s) = \overline{F}(s)$, and with these values as admittances, the circuit right in Fig. 3 has the driving-point admittance $1/\overline{Z}(s) = \overline{F}(s)$. The coefficient identifications are obtained by the explicit formulas on Table 3. The six equivalent circuits representing the impedance $(1-N_m)\overline{F}(s)$ and the six equivalent circuits representing the admittance $(1-N_m)\overline{F}(s)$ are shown in Table 4. The impedance realizations 1M, ..., 6M of $\overline{Z}(s) = F(s)$ and the admittance realizations 7M, ..., 12M of $1/\overline{Z}(s) = F(s)$ are shown in Fig. 4. There are also the duals 1M*, ..., 12M* not shown here. One element in the star corresponding with the branches U,V,W and respectively with the Pi in Fig. 3 is negative. Hence the 1M, ..., 12M and the 1M* ..., 12M* circuits have to be further developed if one wants a realization by positive elements. Some of the circuits allow the application of a perfectly coupled transformer, to others the Bott-Duffin expansion can be applied, and there are some in which the nagative resistor can be realized by a tunnel diode.

The material covered in this paper will be extensively discussed in one of the AFCRL Physical Sciences Papers which is in preparation at present time. It can be obtained with no costs by writing to the author.

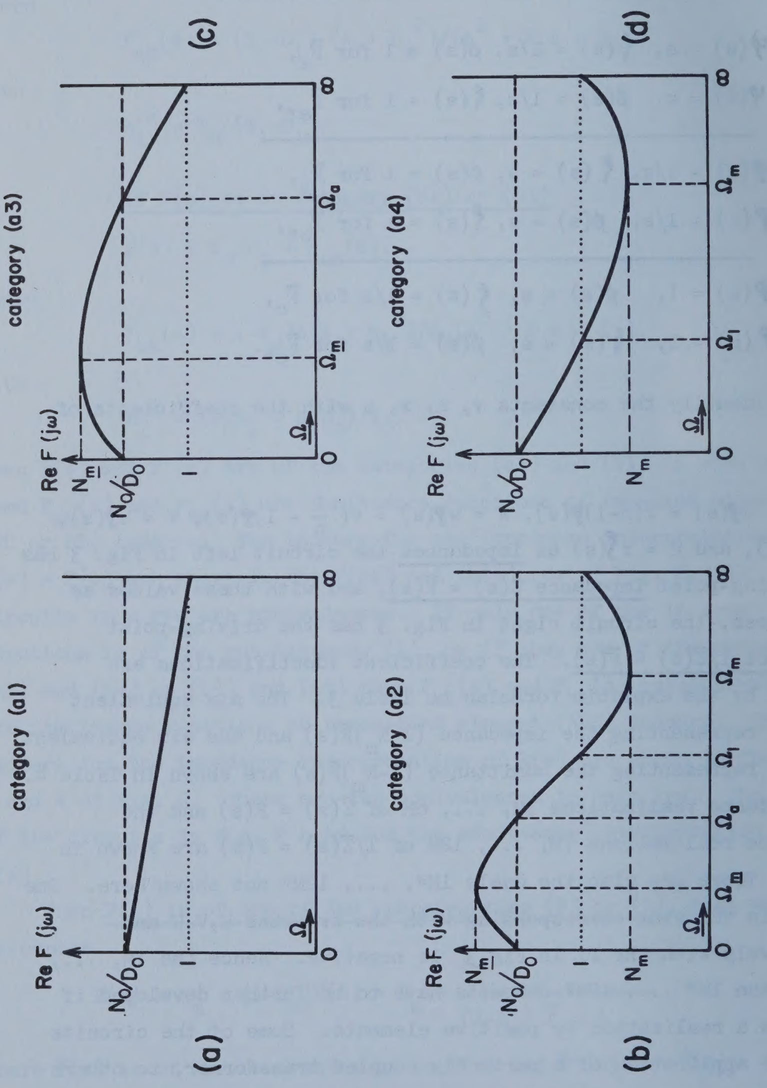


Figure 1

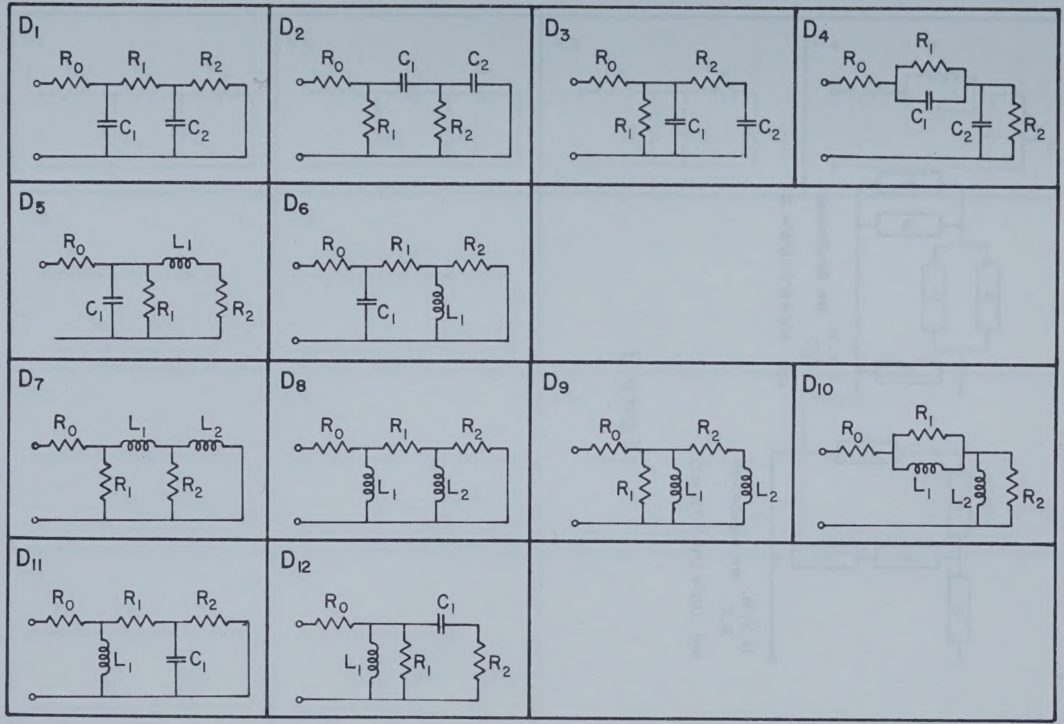
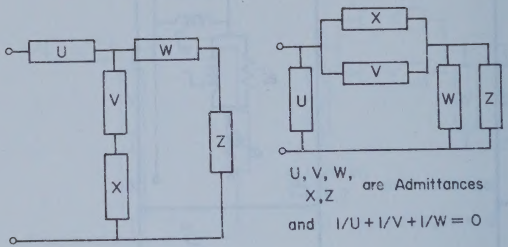

D₁
R₀ R₁ R₂ C₁ C₂
D₂
R₀ C₁ C₂ R₁ R₂
D₃
R₀ R₂ R₁ C₁ C₂
D₄
R₀ R₁ C₁ C₂ R₂
D₅
R₀ L₁ C₁ R₁ R₂
D₆
R₀ R₁ R₂ C₁ L₁
D₇
R₀ L₁ L₂ R₁ R₂
D₈
R₀ R₁ R₂ L₁ L₂
D₉
R₀ R₂ R₁ L₁ L₂
D₁₀
R₀ R₁ L₁ L₂ R₂
D₁₁
R₀ R₁ R₂ L₁ C₁
D₁₂
R₀ C₁ L₁ R₁ R₂

Figure 2


U
W
V
X
Z
X
V
U
W
Z

U, V, W, X,Z are Impedances

and 1/U + 1/V + 1/W = 0

U, V, W, X,Z are Admittances

and 1/U + 1/V + 1/W = 0

Figure 3

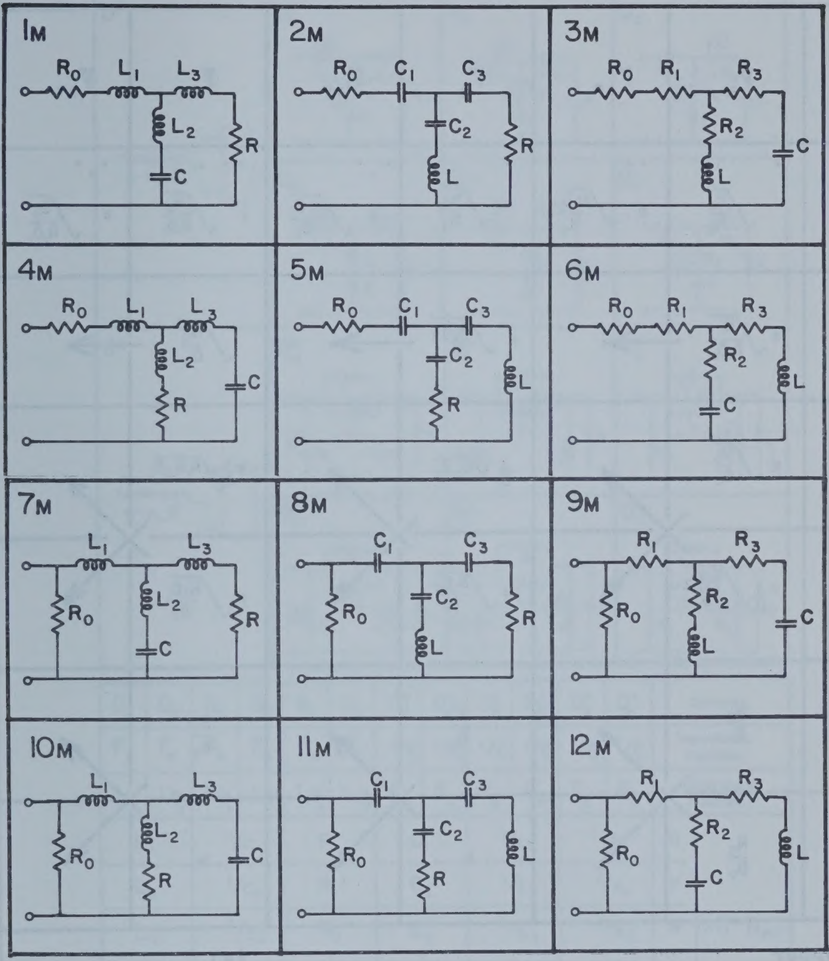
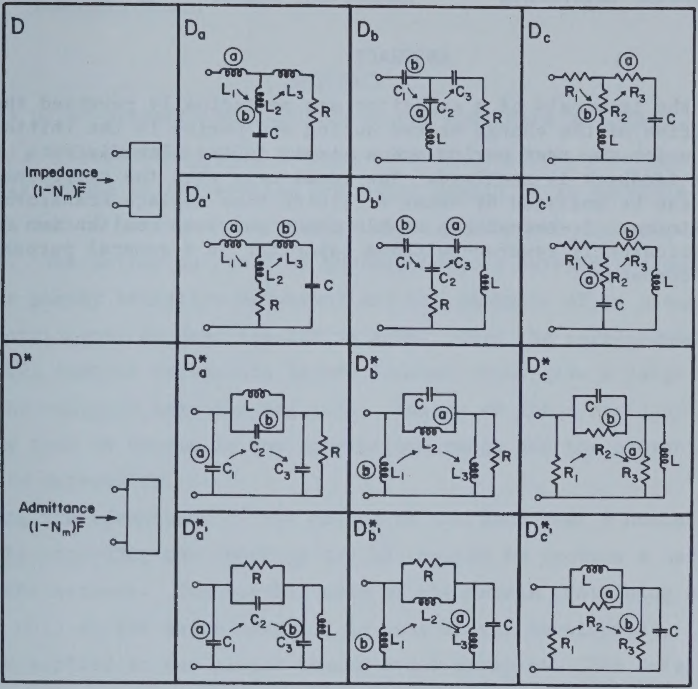

1M
R$_0$
L$_1$
L$_3$
L$_2$
R
C
2M
R$_0$
C$_1$
C$_3$
C$_2$
R
L
3M
R$_0$
R$_1$
R$_3$
R$_2$
C
L
4M
R$_0$
L$_1$
L$_3$
L$_2$
C
R
5M
R$_0$
C$_1$
C$_3$
C$_2$
L
R
6M
R$_0$
R$_1$
R$_3$
R$_2$
L
C
7M
L$_1$
L$_3$
R$_0$
L$_2$
R
C
8M
C$_1$
C$_3$
C$_2$
R$_0$
R
L
9M
R$_1$
R$_3$
R$_0$
R$_2$
C
L
10M
L$_1$
L$_3$
L$_2$
R$_0$
C
R
11M
C$_1$
C$_3$
C$_2$
R$_0$
L
R
12M
R$_1$
R$_3$
R$_2$
R$_0$
L
C

Figure 4

	$\bar{F}_a$	$\bar{F}_{a*}$	$\bar{F}_b$	$\bar{F}_{b*}$	$\bar{F}_c$	$\bar{F}_{c*}$
n	$\cdot$ $\sqrt{\frac{\bar{D}_0}{\bar{N}_0}}$	$\cdot\cdot$ $\pm\sqrt{\frac{\bar{N}_1\bar{D}_1}{\bar{N}_0}}$	$\cdot$ $\sqrt{\frac{\bar{N}_0}{\bar{D}_0}}$	$\because$ $\mp\sqrt{\frac{\bar{N}_1\bar{D}_1}{\bar{D}_0}}$	$\cdot\cdot$ $\pm\sqrt{\frac{\bar{N}_0}{\bar{N}_1\bar{D}_1}}$	$\because$ $\mp\sqrt{\frac{\bar{D}_0}{\bar{N}_1\bar{D}_1}}$
u	$\mp\sqrt{\frac{\bar{N}_1}{\bar{D}_1\bar{D}_0}}$	→	$\pm\sqrt{\frac{\bar{N}_1\bar{N}_0}{\bar{D}_1}}$	→	$\sqrt{\frac{\bar{N}_0}{\bar{D}_0}}$	→
v	$\frac{1}{\bar{D}_1}\sqrt{\frac{\bar{N}_0}{\bar{D}_0}}$	↗	$\frac{1}{\bar{D}_1}\sqrt{\bar{N}_0\bar{D}_0}$	↗	$\pm\frac{1}{\bar{D}_0}\sqrt{\bar{N}_1\bar{N}_0\bar{D}_1}$	↗
w	$\pm\frac{1}{\bar{D}_0}\sqrt{\frac{\bar{N}_1\bar{N}_0}{\bar{D}_1}}$	↘	$\mp\sqrt{\frac{\bar{N}_1\bar{D}_0}{\bar{D}_1}}$	↘	$\mp\sqrt{\frac{\bar{N}_1\bar{D}_1}{\bar{D}_0}}$	↘
x	$\frac{\bar{N}_0}{\bar{D}_1}$	↗	$\frac{1}{\bar{D}_1}$	↗	$\frac{\bar{N}_1}{\bar{D}_0}$	↗
z	$\frac{\bar{N}_0}{\bar{D}_0}$	↘	1	↘	N_1	↘

Upper (a)
signs for n, u, v, w in main category No signs = +
Lower (b)

Table 3


D
Impedance
(I−N_m)F̄
Da
Db
Dc
Da'
Db'
Dc'
D*
Admittance
(I−N_m)F̄
D*a
D*b
D*c
D*a'
D*b'
D*c'
L1
L2
L3
R
C
L
C1
C2
C3
R1
R2
R3
a
b

D_a	$D_{a'}$	D_b	$D_{b'}$	D_c	$D_{c'}$	D^*_a	$D^*_{a'}$	D^*_b	$D^*_{b'}$	D^*_c	$D^*_{c'}$	Circuit	
$\bar{F}_a$	$\bar{F}_{a'}$	$\bar{F}_b$	$\bar{F}_{b'}$	$\bar{F}_c$	$\bar{F}_{c'}$	$1/\bar{F}_a$	$1/\bar{F}_{a'}$	$1/\bar{F}_b$	$1/\bar{F}_{b'}$	$1/\bar{F}_c$	$1/\bar{F}_{c'}$	Impedance Function	
$1/\bar{F}_a$	$1/\bar{F}_{a'}$	$1/\bar{F}_b$	$1/\bar{F}_{b'}$	$1/\bar{F}_c$	$1/\bar{F}_{c'}$	$\bar{F}_a$	$\bar{F}_{a'}$	$\bar{F}_b$	$\bar{F}_{b'}$	$\bar{F}_c$	$\bar{F}_{c'}$	Admittance Function	
L_1		$1/C_1$		R_1		C_1		$1/L_1$		$1/R_1$		u	$\times(1-N_m)$
L_2		$1/C_2$		R_2		C_2		$1/L_2$		$1/R_2$		v	
L_3		$1/C_3$		R_3		C_3		$1/L_3$		$1/R_3$		w	
$1/C$	R	L	R	L	$1/C$	$1/L$	$1/R$	C	$1/R$	C	$1/L$	x	
R	$1/C$	R	L	$1/C$	L	$1/R$	$1/L$	$1/R$	C	$1/L$	C	z	

Table 4

THE PERIODICALLY REVERSE-SWITCHED CAPACITOR

JOHN A. McKINNEY
NAA, Antonetics
Anaheim, California

CHARLES A. HALIJAK
Kansas State University
Manhattan, Kansas

ABSTRACT

If the terminals of a capacitor are periodically reversed then the negative of the charge stored during one period is the initial condition for the next period and a simple device with discrete positive feedback is realized. The ideal case when the switching time is zero can be analyzed by means of finite time Laplace transforms and Z_m-transforms. Z_1-transforms enable ready physical realization of the ideal periodically reverse-switched capacitor on a general purpose digital computer.

AN APPLICATION OF RESIDUAL NETWORKS TO NETWORK SYNTHESIS

L. B. MASTEN AND R. H. SEACAT
Texas Technological College
Lubbock, Texas

ABSTRACT

This paper presents an approach to the synthesis of linear, passive, bilateral, RLC networks using the concepts of residual networks. By using this method in the complex frequency domain it is possible to synthesize a set of networks which will have a prescribed characteristic equation. The method is based on correspondences between residual networks (or purely resistive networks) and the elements of an assumed form of the matrix used to describe the network. Once the correspondences are determined, certain restraints become evident which give a range of values the residual networks may take. Values of these residual networks may then be chosen in the appropriate range and the elements of the matrix determined.

Among the advantages of the method is the fact that a number of restraints regarding the topology can be imposed to produce a desired form of the network. The special case of the matrix containing storage elements only on the major diagonal is used in the development. The method is applied to two simple second order examples. The details of the application can be extended to higher order systems.

INTRODUCTION

Considering the success of the topological techniques as applied to network analysis, it follows that there may be merit in the application of some of these techniques to the problem of synthesis. Classical methods of network synthesis are largely prescriptive methods which outline definite procedures, but an individual approach to each problem is generally required. The result is the production of one solution from a set of possible solutions. Topological approaches, however, may lead to more general forms of synthesis procedures. More of the possible solutions may be determined directly, and several different circuit configurations may be obtained using the same general approach.

Topological synthesis techniques usually consist of the following procedures:

(a) assume certain forms for the branch impedances

(b) determine the minimum number of branches and meshes or nodes necessary to give the desired form of impedances or admittance function

(c) find all of the trees or co-trees of the topology

(d) express the desired function in terms of the trees or co-trees, and

(e) equate the coefficients of the s terms, where s is the complex frequency $\sigma + j\omega$, to determine the element values.

This leads to a very tedious computational problem for even the simplest forms, if the work is not done on a computer. Most of the present topological techniques are variations of this approach.

The method to be considered in this paper is based on the concept of residual networks as developed by Dr. R. H. Seacat. If the network is described by a transfer function $T(s) = \frac{P(s)}{Q(s)}$, where s is the complex variable $\sigma + j\omega$, P(s) and Q(s) are polynomials of s, it has been shown that the coefficients of the s terms can be determined by using the concepts of residual networks. Residual networks are purely resistive networks which remain when the storage elements assume the extreme values of zero and infinity. The individual coefficients are expressed as a sum of terms each of which has a value determined by the values of the storage elements and the determinant of the residual networks. The values of the determinants of the residual networks depend on both the values of the resistive elements and the topology of the network. This fact allows correspondences to be established between the two major parts of topological synthesis, the determination of the network topology and determination of the element values. It will be shown that the Z_{ij} terms of an impedance matrix can be expressed in terms of certain of the residual networks and that the remainder of the residual networks can be determined, thus allowing relations to be established from which the storage element values can be determined.

In considering the problem of synthesis of electrical networks in this paper, it shall be assumed that the questions of economics, tolerances, parasitic elements, stability, and the existence of a solution as a passive network have been resolved and included in the approximation of the function used to describe the network. Further, the network to be realized shall be restricted by requiring that it be comprised of linear, passive, bilateral elements and contain no transformers or dependent sources. For the considerations of this paper, the network is assumed to be described by a driving-point impedance expressed as a ratio of two polynomials involving the complex variable s.

The driving-point impedance of a network may be expressed as,

$$Z_n(s) = \frac{\Delta(s)}{\Delta_{nn}(s)} , \tag{1}$$

where $Z_n(s)$ is the driving-point impedance as seen by a voltage source in mesh n. $\Delta(s)$ is the determinant of the network on a mesh basis. $\Delta_{nn}(s)$ is the determinant obtained by collapsing row n and column n of the original network determinant. If these two determinants are expanded, the impedance, Z(s), may be expressed in the form:

$$Z(s) = \frac{a_n(s)^n + a_{n-1}s^{n-1} + \dots + a_n}{b_m s^m + b_{m-1}s^{m-1} + \dots + b_0} \tag{2}$$

Thus (s) is defined as:

$$\Delta(s) = \frac{a_n s^n + a_{n-1}s^{n-1} + \dots + a_0}{Ks^P} \tag{3}$$

And $\Delta_{nn}(s)$ will have the form:

$$\Delta_{nn}(s) = \frac{b_m s^m + b_{m-1}s^{m-1} + \dots + b_0}{Ks^P} \tag{4}$$

where K and P of Equations 3 and 4 are the same.

A. The Numerator.

Consider the following expansion of the numerator determinant about Z_1,

$$\Delta = Z_1\Delta_1 + \Delta_{1'}, \tag{5}$$

where Z_1 represents only the storage elements contained in branch 1. If the principal determinant is expanded about Z_2, the Equation 5 will be expanded as:

$$\Delta = Z_1Z_1\Delta_{12} + Z_1\Delta_{12'} + Z_2\Delta_{1'2} + \Delta_{1'2'}. \tag{6}$$

If there are three branches containing storage elements in the network, the resulting expansion would be:

$$\Delta = Z_1Z_2Z_3\Delta_{123} + Z_1Z_2\Delta_{123'} + Z_1Z_3\Delta_{12'3} + Z_2Z_3\Delta_{1'23} + Z_1\Delta_{12'3'} + Z_2\Delta_{1'23'} + Z_3\Delta_{1'2'3} + \Delta_{1'2'3'}. \tag{7}$$

In Equations 5, 6, and 7, the primed subscripts refer to impedances indicated by that subscript being set equal to zero while the unprimed subscripts refer to impedances referred to by that subscript being set equal to infinity. The subscripted delta terms refer to the determinants of the resistive network remaining when the original network has been subjected to the restrictions indicated by the subscripts. The networks corresponding to each of these delta terms are called residual networks.

Suppose that Z_1 is an inductance while Z_2 and Z_3 are capacitances, then Equation 7 becomes:

$$\Delta = \frac{sL_1}{s^2 C_2 C_3} \Delta_{123} + \frac{sL_1}{sC_2} \Delta_{123'} + \frac{sL_1}{sC_3} \Delta_{12'3} + \frac{\Delta_{1'23}}{s^2 C_2 C_3}$$

$$+ sL_1 \Delta_{12'3'} + \frac{\Delta_{1'23'}}{sC_2} + \frac{\Delta_{1'2'3}}{sC_3} + \Delta_{1'2'3'}. \qquad (8)$$

This leads to the following expansion for the principal determinant:

$$\Delta = \frac{s^3 L_1 C_2 C_3 \Delta_{12'3'} + s^2 (L_1 C_2 \Delta_{123'} + C_2 C_3 \Delta_{1'2'3'} + L_1 C_2 \Delta_{12'3})}{s^2 C_2 C_3}$$

$$+ \frac{s(L_1 \Delta_{123} + C_2 \Delta_{1'2'3} + C_3 \Delta_{1'23'}) + \Delta_{1'23}}{s^2 C_2 C_3} \qquad (9)$$

For a more consistant form let the subscripts refer to the element values (L's and C's) rather than the impedances, with the primes having the same connotation as before except now applied to the element values rather than the impedance. Equation 9 becomes:

$$\Delta = \frac{s^3 L_1 C_2 C_3 \Delta_{123} + s^2 (L_1 C_2 \Delta_{123'} + L_1 C_3 \Delta_{12'3} + C_2 C_3 \Delta_{1'23})}{s^2 C_2 C_3}$$

$$+ \frac{s\,(L_1 \Delta_{12'3'} + C_2 \Delta_{1'23'} + C_3 \Delta_{1'2'3}) + \Delta_{1'2'3'}}{s^2 C_2 C_3}. \qquad (10)$$

This could be generalized by saying that the coefficients of s^K for an n^{th} order polynomial is the sum of the products of each combination of the n elements taken K at a time, multiplied by a delta having all of the subscripts primed which do not appear in the combination of elements. It is noted that the denominator of Equation 10 is the product of all capacitances contained in the determinant. The coefficients thus defined can be related to the coefficients of Equations 3 or 4 as the case may be. In considering open and short-circuit conditions, it is noted that the primed inductances are short-circuits while the primed capacitances are open-circuits.

Consider a network containing n storage elements as indicated in Figure 1. The mesh impedance matrix for the network of Figure 1 could have the form:

$$(Z) = \begin{bmatrix} (R_{11} + sL_1) & R_{12}\ldots & & R_{1n} \\ R_{21} & (R_{22} + sL_2) & \ldots & R_{2n} \\ R_{m1}\ldots & & (R_{mm} + \frac{1}{sC_m})\ldots & R_{mn} \\ R_{n1}\ldots & & R_{nm} & (R_{nn} + \frac{1}{sC_n}) \end{bmatrix} \tag{11}$$

It is observed that if all the storage elements are replaced by open circuits the determinant formed from the impedance matrix will be unity, zero, or a positive real value depending on the topology of the network. If the value of the determinant is unity when all storage elements are open-circuited this implies that the resulting graph is connected and contains no closed paths. If the determinant has a positive real value this implies that the graph resulting from open-circuiting the storage elements is connected but does contain a closed resistive path. For the present consider that these resistances are combined in some manner to give no closed paths and the determinant of the resulting graph under open-circuit conditions will be unity. This is denoted by:

$$\Delta_{123\ldots(m-1)m'\ldots n'} = 1 \tag{12}$$

where the first (m-1) subscripts refer to inductances which are infinite or open-circuited when unprimed. Subscripts m' through n' refer to primed or zero capacitances which are also open circuits.

In this case if all storage elements are opened except one, the resulting determinant will be of the form:

$$R_{ii} = \Delta_{12\ldots i\ldots(m-1)m'\ldots n'} \tag{13}$$

if i is the subscript of an inductance or

$$R_{ii} = \Delta_{12\ldots(m-1)m'\ldots i\ldots n'} \tag{14}$$

if i is the subscript of a capacitance.

Let all of the storage elements be opened except i and j. The determinant of the impedance matrix then becomes:

$$\Delta_{12\ldots i'\ldots j'\ldots(m-1)(m'\ldots n')} = \begin{vmatrix} R_{ii} & R_{ij} \\ R_{ij} & R_{jj} \end{vmatrix} = R_{ii}R_{jj} - R_{ij}^{2} \tag{15}$$

where i and j will be primed if they refer to inductances or unprimed if they refer to capacitances. From Equation 15 it is seen that

$$R_{ij}^{2} = R_{ii}R_{ij} - \Delta_{12\ldots i'\ldots j'\ldots(m-1)m'\ldots n'} \tag{16}$$

Using Equations 13 or 14 in Equation 16 leads to:

$$R_{ij} = [(\Delta_{12\ldots i'\ldots(m-1)m'\ldots n'})(\Delta_{12\ldots j'\ldots(m-1)m'\ldots n'}) - \Delta_{12\ldots i'\ldots j'\ldots(m-1)m'\ldots n'}]^{1/2} \tag{17}$$

For a circuit containing three storage elements L_1, L_2, and C_3, Equations 13 and 17 become:

$$\begin{aligned}
R_{11} &= \Delta_{1'23'}, \\
R_{22} &= \Delta_{12'3'}, \\
R_{33} &= \Delta_{123}, \\
R_{12} &= [\Delta_{1'23'}\Delta_{12'3'} - \Delta_{1'2'3'}]^{1/2} \\
R_{13} &= [\Delta_{1'23'}\Delta_{123} - \Delta_{1'23}]^{1/2} \\
R_{23} &= [\Delta_{12'3'}\Delta_{123} - \Delta_{12'3}]^{1/2}
\end{aligned} \tag{18}$$

From Equation 17 it is observed that in order for R_{ij} to be real the following inequality must hold:

$$(\Delta_{12\ldots i'\ldots(m-1)m'\ldots n'})(\Delta_{12\ldots j'\ldots(m-1)m'\ldots n'}) \geq \Delta_{12\ldots i'\ldots j'\ldots(m-1)m'\ldots n'}. \tag{19}$$

Other restrictions are possible, for example, if the circuit is mappable and the equations are written on a mesh basis then,

$$R_{mm} \geq R_{m1} + R_{m2} + \ldots + R_{m(m-1)} + R_{m(m+1)} + \ldots + R_{mn}. \tag{20}$$

A further restriction on the values of the residual networks is obtained by observing that, if the coefficients of Equation 3 are equated to the coefficients of the corresponding expansion involving residual networks, there are n + 1 equations and n storage elements. The last term in the expansion involving residual networks does not involve storage elements. Thus there are n equations and n storage elements which can be considered. The values of the storage elements may be determined by solution of the n equations. It is necessary that the residual networks be chosen so that the solution of the equations yield positive real values for the storage elements.

It is important to note that if all of the $\Delta_{123\ldots i'\ldots(m-1)m'\ldots n'}$ terms and all of the $\Delta_{123\ldots i'\ldots j'\ldots(m-1)m'\ldots n'}$ terms are known, then from Equations 13 and 17 all of the R_{ij} terms can be determined. Also, it is noted that when the R_{ij} are known, the remaining residual networks may be determined. When the network determinant is expanded using the interpretation derived from residual networks, n equations involving the known residual networks and the unknown storage elements may be solved for the storage element values, thus completing the synthesis of the function.

EXAMPLE

A. An RL Function.

Consider the expanded determinant of an RL network given in the form:

$$\Delta(s) = a_2 s^2 + a_1 s + a_0. \tag{21}$$

The expansion in terms of the storage elements and residual networks will be in the form:

$$\Delta(s) = L_1 L_2 \Delta_{12} s^2 + (L_1 \Delta_{12'} + L_2 \Delta_{1'2}) s + \Delta_{1'2'}. \tag{22}$$

If the coefficients of Equations 21 and 22 are equated the resulting equations will be:

$$L_1 L_2 \Delta_{12} = a_2, \tag{23a}$$

$$L_1 \Delta_{12'} + L_2 \Delta_{1'2} = a_1, \tag{23b}$$

$$\Delta_{1'2'} = a_0. \tag{23c}$$

Solving Equations 23a and 23b for L_1 and L_2 gives

$$L_1 = \frac{1}{\Delta_{12'}} \left[\frac{a_1}{2} \pm \left(\left(\frac{a_1}{2}\right)^2 - \frac{a_1 \Delta_{12'} \Delta_{1'2}}{\Delta_{12}}\right)^{1/2}\right], \tag{24a}$$

$$L_2 = \frac{a_2}{L_1 \Delta_{12}} \tag{24b}$$

If it is necessary for L_1 to be real then:

$$\left(\frac{a_1}{2}\right)^2 \geq \left(\frac{a_2 \Delta_{12'} \Delta_{1'2}}{\Delta_{12}}\right). \tag{25}$$

Equations 13 and 17 become respectively:

$$R_{11} = \Delta_{1'2}, \tag{26a}$$

$$R_{22} = \Delta_{12'}, \tag{26b}$$

$$R_{12} = (\Delta_{1'2} \Delta_{12'} - \Delta_{1'2'})^{1/2} \tag{26c}$$

Inequality (19) becomes:

$$\Delta_{1'2} \Delta_{12'} \geq \Delta_{1'2'}. \tag{27}$$

Combining inequalities (25) and (27) gives:

$$\frac{a_1^2}{4a_2} \Delta_{12} \geq \Delta_{12'}\Delta_{1'2} \geq \Delta_{1'2'}. \tag{28}$$

If Δ_{12} is assumed to be unity which implies that all current paths are interrupted but leave a connected graph, and using Equation 23c it follows that:

$$\frac{a_1^2}{4a_2} \geq \Delta_{12'}\Delta_{1'2} \geq a_0. \tag{29}$$

In particular suppose that $\Delta(s)$ has the form:

$$\Delta(s) = 2s^2 + 5s + 2. \tag{30}$$

Then inequality (29) becomes:

$$\frac{\Delta_{25}}{4(2)} \geq \Delta_{12'}\Delta_{1'2} \geq 2 \tag{31}$$

If values of $\Delta_{12'}$ and $\Delta_{1'2}$ are chosen so that inequality (31) is satisfied then R_{11}, R_{22}, and R_{12} may be determined from Equations 26 and L_1 and L_2 may be determined from Equation 24. Figure 2 shows several of the networks that result from the choice of $\Delta_{1'2}$ and $\Delta_{12'}$ within the limits of Equation 31.

If an impedance function is to be realized, the procedure is the same except that there are additional restrictions imposed. For example:

$$Z(s) = \frac{a_2 s^2 + a_1 s + a_0}{b_1 s + b_0}. \tag{32}$$

Z(s) could also be written as:

$$Z(s) = \frac{s^2 L_1 L_2 \Delta_{12} + s(L_1\Delta_{12'} + L_2\Delta_{1'2}) + \Delta_{1'2'}}{sL_2\Delta_{12} + \Delta_{12'}}$$

Equating coefficients of Equation 32 and 33 gives:

$$L_1 L_2 \Delta_{12} = a_2,$$

$$L_1\Delta_{12'} + L_2\Delta_{1'2} = a_1,$$

$$\Delta_{1'2'} = a_0,$$

$$L_2\Delta_{12} = b_1,$$

$$\Delta_{12'} = b_0. \tag{33}$$

Equation 33 can be solved for L_1, L_2, $\Delta_{1'2}$, and $\Delta_{12'}$ from which the values of R_{11}, R_{22}, and R_{12} can be determined using

Equations 24 and 26.

CONCLUSIONS

The method of synthesis described in this paper is dependent on the determination of the values of the residual networks. That is, if the residual network values are known, the storage element values may be determined by solving a set of equations obtained by equating coefficients. The problem is then primarily one of determining the residual network values.

The values of the determinant of the residual networks depend on both the values of the individual resistances and the topology of the network. One set of restrictions on the values of the determinants of the residual networks can be obtained since the solutions for the values of the storage elements must be positive real values. At this point there are at least two possible methods of obtaining further restrictions. One method would be to specify the desired form of the topology of the network and proceed as was done in the paper. Another method that may have merit would be to choose the values of the residual networks consistant with the restrictions that the storage elements must be positive and real and then to synthesize a resistive network that would give the desired determinant values under the various open and short-circuit conditions.

It would also be possible to choose the values of the storage elements and then determine the values of the determinants of the residual networks to give the desired function. The remaining problem would be to synthesize a resistive network that would give the desired residual network values.

The concept of residual networks as applied to synthesis gives a flexibility that could be very useful. We feel that this approach to the synthesis problem merits further study.

REFERENCES

1. E. D. Merkl and R. H. Seacat, "Topological Methods Applied to Active Networks," Eighth Midwest Symposium on Circuit Theory, Fort Collins, Colorado, June, 1965.

2. R. H. Seacat, "A Method of Network Analysis Using Residual Networks," Ph.D. Dissertation, Texas A&M University, College Station, Texas, 1963.

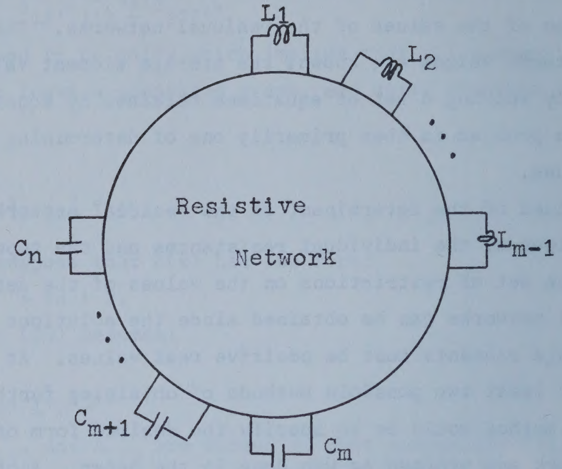

L1
L2
Resistive
Network
Cn
Lm-1
Cm+1
Cm

Figure 1. Network Containing n Storage Elements

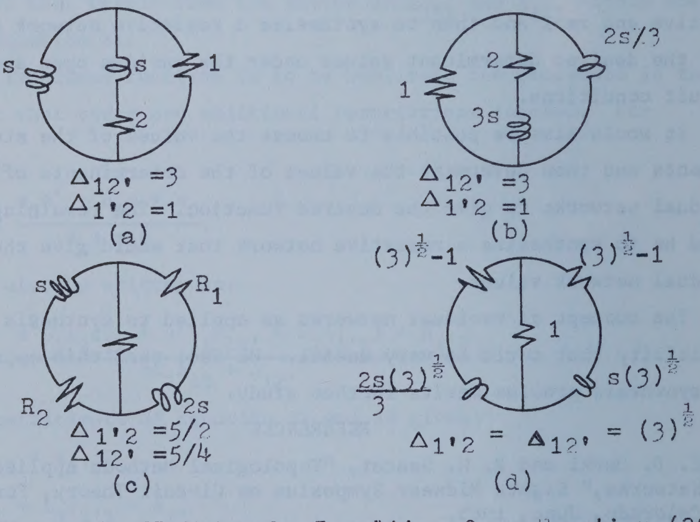

s
s
1
2
$\Delta_{12'} = 3$
$\Delta_{1'2} = 1$
(a)
2s/3
2
1
3s
$\Delta_{12'} = 3$
$\Delta_{1'2} = 1$
(b)
s
R1
R2
2s
$\Delta_{1'2} = 5/2$
$\Delta_{12'} = 5/4$
(c)
$(3)^{\frac{1}{2}}-1$
$(3)^{\frac{1}{2}}-1$
1
$\frac{2s(3)^{\frac{1}{2}}}{3}$
$s(3)^{\frac{1}{2}}$
$\Delta_{1'2} = \Delta_{12'} = (3)^{\frac{1}{2}}$
(d)

Figure 2. RL Networks Resulting from Equation (30)

THE APPLICATION OF SIGNAL FLOW GRAPH ANALYSIS TO A NETWORK HAVING A TIME VARYING ELEMENT

ROLLY HASSUN
Hewlett-Packard Company
Palo Alto, California

ABSTRACT

The techniques for analyzing electrical networks fall into two main categories. The first consists in immediate translation of the given physical information into mathematical language and processing it in the "mathematical domain" to yield useful relations about the network. This method alienates one from the physical model and thereby limits one's insight into the physical phenomenon. The second category, which is well typified by analytical techniques such as signal flow graph analysis, allows one to maintain contact with the physical model. As a result, one's understanding of the functioning of the physical model is greatly enhanced.

In the analysis of linear networks with time varying elements, it has been normal practice thus far to use techniques belonging to the first category. This paper describes a method for applying signal flow graph analysis to linear networks with time varying elements.

The analysis of a resonant circuit with a time harmonic capacitor is a very good example and the flow graph shows that it is a combination of two types of parametric amplifiers, the upper sideband upconverter and the negative resistance amplifier.

TOPOLOGICALLY DISTINCT SUBNETS IN A COMPLEX NETWORK

HARTLEY JENSEN
University of California
Davis, California

J. W. LA PATRA
University of California
Davis, California

ABSTRACT

There are many things one might want to know about a complex network. The linear graph representation of such a network has provided a fruitful avenue to a variety of useful results. Invariably the results have been associated with some aspect of a subgraph of the basic structure. The subgraph may have been a path, a tree, a circuit, or some group of network elements chosen according to a set of defining criteria. This paper deals with the generalized class of subnets in a network by considering the following topics:

1. Utilization of the transition matrix to generate topologically distinct subgraph classes.

2. Identification of topologically identical vertices in a subgraph class.

3. Determination of the number of subgraphs in each class.

4. Criteria for constraining the generalized subgraph to a particular form.

BACKGROUND

Consideration of some of the most recent papers (references 1, 2, and 3) on the subject of enumeration of topological quantities in a linear graph reveals some interesting limitations. The primary limitation has been due to excessive concern about generating trees for use in a computer analysis of a network. The intense interest in this effort has obscured the natural trade-off which will occur if a more general approach to enumerating many topological quantities is made. The four "variables" to be reasonably balanced are:

1. Number of different topological entities to be generated.

2. Amount of redundant generation to be permitted.

3. Minimization of computer time and storage required.

4. Ease and flexibility of the method of generation from the viewpoint of the user.

In one way or another, most references on enumeration do not satisfy all four of the requirements.

The authors required in their work with graphs, a flexible, general method of enumeration which offered a reasonable balance of the four variables previously listed. The paper presents the results of finding such a method. Its main features are:

1. Many different quantities may be generated.

2. Redundant generation is minimized.

3. Computer time and storage requirements are reasonable for modest vertex complexities.

4. The method is conceptually satisfying, and is easy to program for nearly any application visualized.

INTRODUCTION

The purpose of this paper is to present an approach which uses the theorems of graph theory, the concept of permutations, and the viewpoint of abstract isomorphism to make practical the application of the digital computer to the classification and enumeration problems of networks. The networks will be represented as non-oriented incomplete graphs. The basic question to be answered is: How many topologically distinct ways are there to remove a given number of elements from a complete graph? By topologically distinct one should infer that each removal technique generates a set of isomorphic graphs. Isomorphic graphs are those in which there exists a one to one correspondence between elements and vertices irrespective of how they are designated for identification.

When this question is answered, a means for classifying the ways become available. Those ways which are topologically equivalent will be treated as a class. Determining which class the physical problem falls into eliminates the need to examine those possibilities which are contained in the other classes.

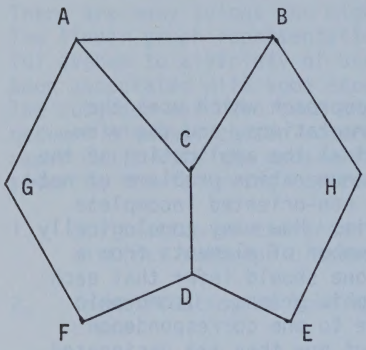
The consideration which follows is that of enumerating the remaining possibilities. This is treated as a problem of enumerating the number of ways the class can be occupied, followed by a listing of the ways the naming of elements and vertices may be permuted for each abstract way of occupying the class.

TRANSITION MATRIX

It requires only brief experience with linear graphs to realize that the way a graph is drawn may provide insight to topological problems. Complex graphs which are identical may not be so recognized. The transition matrix can be used to display symmetry in a manner independent of the way the graph is drawn.

For the limited purpose of treating non-oriented linear graphs where no more than one element (edge) may be found connecting any two vertices, the transition matrix[4] will be defined as follows: The rows and columns are to be named after the vertices. It will further be required that the 'i th' row and the 'i th' column be named after the same vertex. At the intersection of a given row and column an entry is made which may either denote the name of the element connecting the two vertices or may simply state the existence or absence of such a connection.

The symmetry of the transition matrix provides an easy means to determine if a reordering of the vertex designations will display the isomorphism (congruency) of two incomplete graphs. In the following figure, removal of element AC is seen to be no different from removing BC.



8 Vertex graph

Figure 1a

	A	B	C	D	E	F	G	H
A	0	1	1	0	0	0	1	0
B	1	0	1	0	0	0	0	1
C	1	1	0	1	0	0	0	0
D	0	0	1	0	1	1	0	0
E	0	0	0	1	0	0	0	1
F	0	0	0	1	0	0	1	0
G	1	0	0	0	0	1	0	0
H	0	1	0	0	1	0	0	0

Transition Matrix with Entries Indicating Existence of Elements

Figure 1b

	A	B	C	D	E	F	G	H
A	0	1	X	0	0	0	1	0
B	1	0	1	0	0	0	0	1
C	X	1	0	1	0	0	0	0
D	0	0	1	0	1	1	0	0
E	0	0	0	1	0	0	0	1
F	0	0	0	1	0	0	1	0
G	1	0	0	0	0	1	0	0
H	0	1	0	0	1	0	0	0

Element AC Removed

Figure 2a

	A	B	C	D	E	F	G	H
A	0	1	1	0	0	0	1	0
B	1	0	X	0	0	0	0	1
C	1	X	0	1	0	0	0	0
D	0	0	1	0	1	1	0	0
E	0	0	0	1	0	0	0	1
F	0	0	0	1	0	0	1	0
G	1	0	0	0	0	1	0	0
H	0	1	0	0	1	0	0	0

Element BC Removed

Figure 2b

	B	A	C	D	F	E	H	G
B	0	1	X	0	0	0	1	0
A	1	0	1	0	0	0	0	1
C	X	1	0	1	0	0	0	0
D	0	0	1	0	1	1	0	0
F	0	0	0	1	0	0	0	1
E	0	0	0	1	0	0	1	0
H	1	0	0	0	0	1	0	0
G	0	1	0	0	1	0	0	0

Fig 2b Re-ordered

Figure 2c

In this example there are 8! or 40,320 ways the elements may be ordered, and it will be difficult to determine significant orderings by trial and error. For 50 vertices which may be found in a communications network, the 50! possibilities allow one to say that trial and error is virtually impossible. This example leads to the following question: How many ways can one remove a given number of elements from a complete graph and form sets of incomplete graphs which are topologically identical?

In a 5 vertex complete graph there are 120 ways to remove three elements. By inspection, there are only four topologically distinct ways to remove the three elements. These four ways are shown in Figure 3 where the dotted lines represent missing elements.

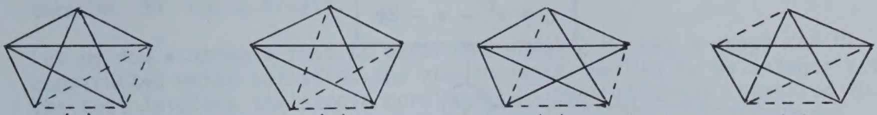

(a) (b) (c) (d)

Figure 3

The validity of the four distinct forms may be verified by noting that the subgraphs formed by removing any one of the 120 possible combinations is isomorphic to one of the above figures. A method is required to handle problems which are so complex the result cannot be obtained by inspection.

POSSIBLE INCOMPLETE GRAPHS

There are $(v^2 - v)$ entries in the transition matrix for a complete non-oriented graph of v vertices. When an element is removed from the graph two entries are eliminated from the transition matrix. The sum of the summation of the entries in the rows or columns is

$$c = v^2 - v - 2e$$

where e is the number of elements removed. The sum of either the row or column entries is also the totality of the vertex degrees.

A general classification of a graph according to vertex order will now be developed. The following convention will be adopted. Given a v vertex graph, let the vertices be labeled $v_1, v_2, \ldots, v_n$, with degrees $d_1, d_2, \ldots, d_n$, respectively.

Without losing generality, a convention may be adopted which requires the vertices to appear in ascending degree according to suffix, i.e., that

$$d_1 \leq d_2 \leq \ldots\ldots \leq d_n$$

It is now possible, with the aid of a small digital computer, to find all the compositions[5] of the number $c = v^2 - v - 2e$ which represent the possible incomplete graphs formed from the "v" vertex complete graph with "e" elements missing.

The ordered combinations are generated with a nested Fortran "Do" loop. A flow chart is shown in Figure 4.

A 5 vertex graph with three elements missing requires the summation of row sums to be 14. One possibility and its transition matrix are shown in Figures 5a and 5b.

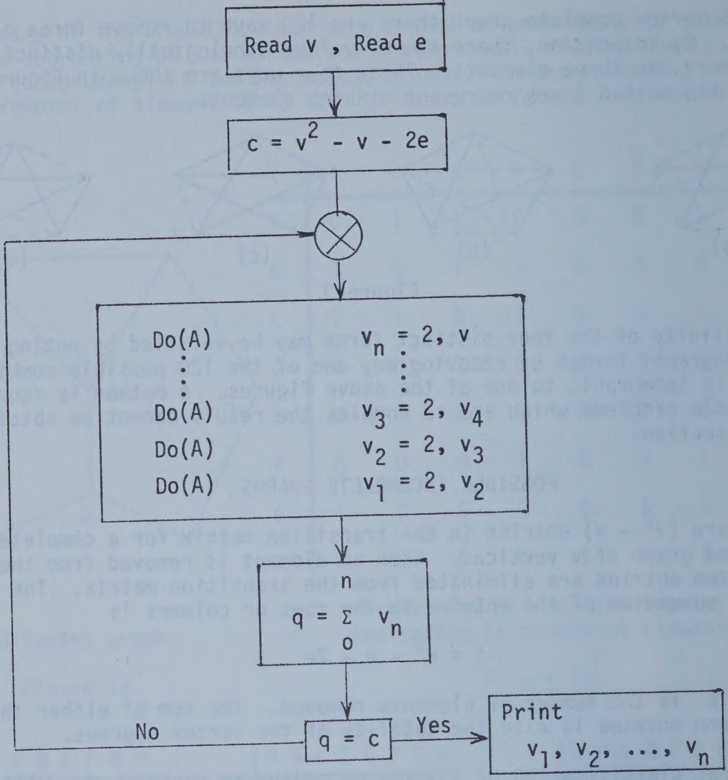

Read v , Read e
$c = v^2 - v - 2e$
Do(A) $v_n = 2, v$
Do(A) $v_3 = 2, v_4$
Do(A) $v_2 = 2, v_3$
Do(A) $v_1 = 2, v_2$
$q = \sum_{o}^{n} v_n$
$q = c$
No
Yes
Print $v_1, v_2, \ldots, v_n$

Figure 4

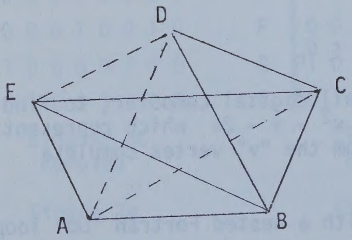


Figure 5a

	A	B	C	D	E		
A	0	1	0	0	1	=	2
B	1	0	1	1	1	=	4
C	0	1	0	1	1	=	3
D	0	1	1	0	0	=	2
E	1	1	1	0	0	=	3
							14

Figure 5b

The total of the row sums is equal to the sum of the vertex degrees. The composition of the number 14 into five numbers between 2 and 5 yields the following possible row sums:

a.	22244	d.	23333
b.	13334	e.	11444
c.	22334	f.	12344

Note that a. through d. correspond to the graphs of Figure 3a through 3d respectively.

Let us now examine a test to determine whether or not a graph can be constructed which satisfies the requirements implied by an integer set. The test involves the simple consideration that elements connected to one vertex must also be connected to another vertex. If one cannot account for all possible connections, the graph cannot be constructed.

Step 1 - Construct an empty v x v matrix.

Step 2 - Identify the rows and columns with the vertex degree so that the degree of the vertices will appear in ascending order.

Step 3 - Enter a zero in all locations along the diagonal.

Step 4 - Enter the highest degree vertex entries in the v th row and v th column in as many ways as are possible.

Step 5 - For each way of Step 4 make entries of the next vertex in the (v-1)th row and (v-1)th column in as many ways as possible.

Step 6 - Continue with each new row until it is determined either:

1. An entry cannot be made without violating the vertex degree in which case the graph may not exist.
2. All entries are made in a manner which satisfies the vertex degree requirements in which case the graph definitely exists.

Step 7 - If in Step 6, result 1 is obtained, the remaining possibilities are examined to exhaustion unless at some point result 2 is obtained. If the possibilities are exhausted without obtaining result 2, then it may be said that the graph definitely does not exist. Examples of compositions for which graphs do not exist are 11444 and 12344 from the previous 5 vertex example.

An example which is worked out by hand follows, but it should be emphasized that the digital computer would normally do the testing. The reader should also note that the number of possibilities to be tested is not as large as one might guess at first glance. This is a result of the requirement that entries be made in both rows and columns which reduces the freedom of choice for each succeeding vertex. Therefore, the test converges very rapidly.

12344

Steps 1 and 2

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 4 \\ \hline 1 & & & & & \\ 2 & & & & & \\ 3 & & & & & \\ 4 & & & & & \\ 4 & & & & & \end{array}$$

Step 3

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 4 \\ \hline 1 & 0 & & & & \\ 2 & & 0 & & & \\ 3 & & & 0 & & \\ 4 & & & & 0 & \\ 4 & & & & & 0 \end{array}$$

Step 4

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 4 \\ \hline 1 & 0 & & & & 1 \\ 2 & & 0 & & & 1 \\ 3 & & & 0 & & 1 \\ 4 & & & & 0 & 1 \\ 4 & 1 & 1 & 1 & 1 & 0 \end{array}$$

only one way to enter 5
elements in row 5 and
column 5

Step 5

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 4 \\ \hline 1 & 0 & & & 1 & 1 \\ 2 & & 0 & & 1 & 1 \\ 3 & & & 0 & 1 & 1 \\ 4 & 1 & 1 & 1 & 0 & 1 \\ 4 & 1 & 1 & 1 & 1 & 0 \end{array}$$

only one way to enter 5
elements in row 4 and
column 4

Figure 6

Step 6 - Step 5 violated an established vertex degree by requiring a row 1 summation of 2, therefore, this composition cannot exist.

The test of the composition 11444 would have proceeded to the same conclusion in an identical manner. One would, therefore, program the computer to discard all combinations having a 1 in the first location, and 4 in the last two locations.

In the 5 vertex case it was only coincidental that the composition of the number 14 produced all topologically different possibilities. In general, it is not that simple, as the following example will show. Consider a six vertex graph with 10 elements removed. There are $\frac{(6)(6-1)}{2} - 10 = 5$ elements left. (These are the trees of the 6 vertex graph.)

The possible compositions are

112222
111223
111133
111124
111115

Graphs corresponding to these possibilities are:

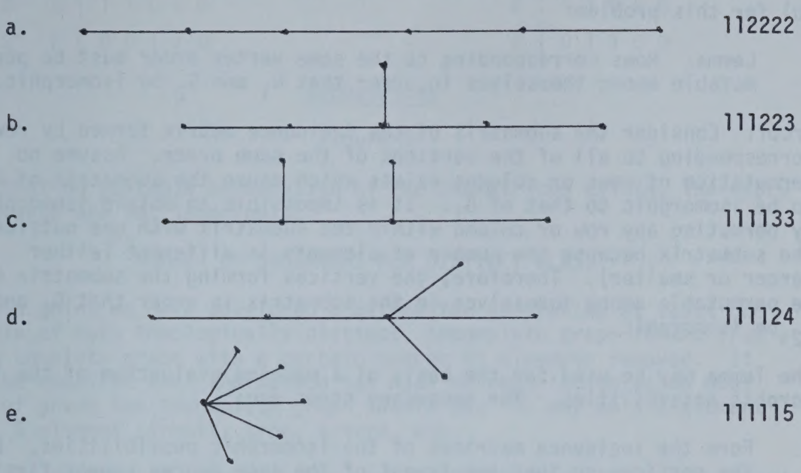


Figure 7

It is not possible to construct other topologically different graphs satisfying the conditions of Figure 7 (a,c,d,e). This is not the case for 111223 of Figure 7b.

The graph below has the required vertex sum requirement but is topologically different, that is, the transition matrices for the two are not isomorphic.

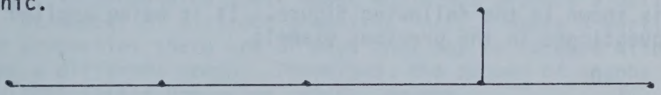


Figure 8

A well known theorem states that if two graphs have incidence matrices which differ only by a permutation of rows and columns, then the graphs are isomorphic[4]. The incidence matrix is related to the transition matrix by:

$$C = A+ W (A^-)^T$$

where C = transition matrix

W = incidence matrix

$A+$ = positive incidence matrix

A^- = negative incidence matrix

Thus, the transition matrix may be thought of as a deformed and rotated space of the incidence matrix. The incidence matrix will be used with the understanding that the same thing could have been accomplished in the transition matrix at the expense of more difficult manipulation.

The following lemma is derived from the theorem and is more directly useful for this problem:

> Lemma: Rows corresponding to the same vertex order must be permutable among themselves in order that G_1 and G_2 be isomorphic.

Proof: Consider the submatrix of the incidence matrix formed by rows corresponding to all of the vertices of the same order. Assume no permutation of rows or columns exists which cause the submatrix of G_2 to be isomorphic to that of G_1. It is impossible to obtain isomorphism by permuting any row or column within the submatrix with one outside the submatrix because the number of elements is different (either larger or smaller). Therefore, the vertices forming the submatrix must be permutable among themselves in the submatrix in order that G_1 and G_2 be isomorphic.

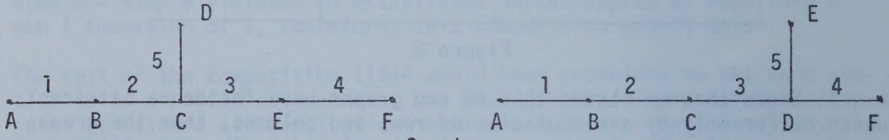
The lemma may be used for the basis of a machine evaluation of the isomorphic possibilities. The necessary steps are:

1. Form the incidence matrices of the isomorphic possibilities. List the vertices so that the fewest of the same degree appear first and so on.

2. Partition the incidence matrices to isolate the submatrices formed by rows corresponding to all the vertices of the same order.

3. Examine the rows and columns of the submatrices to determine if any reordering of the element or vertex designations will display a submatrix identity and therefore isomorphism.

This test is shown in the following figure. It is being applied to the two trees questioned in the previous example.



Graphs

Element

		1	2	3	4	5
3	C	0	1	1	0	1
2	B	1	1	0	0	0
2	E	0	0	1	1	0
1	A	1	0	0	0	0
1	D	0	0	0	0	1
1	F	0	0	0	1	0

	1	2	3	4	5
D	0	0	1	1	1
B	1	1	0	0	0
C	0	1	1	0	0
A	1	0	0	0	0
E	0	0	0	0	1
F	0	0	0	1	0

Incidence Matrices

$$\begin{array}{cc|ccccc} & & 1 & 2 & 3 & 4 & 5 \\ \hline 2 & B & 1 & 1 & 0 & 0 & 0 \\ 2 & E & 0 & 0 & 1 & 1 & 0 \end{array} \qquad \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline B & 1 & 1 & 0 & 0 & 0 \\ C & 0 & 1 & 1 & 0 & 0 \end{array}$$

Submatrices

Figure 9

Note that no suitable permutation of the submatrices exists, therefore the graphs are not isomorphic.

NUMBER OF SUBGRAPHS PER DISTINCT CLASS

At this point we have developed a method for generating at least one example of each topologically distinct, incomplete graph formed from a given complete graph with a certain number of elements removed. It will be recalled that the method has made no restriction as to what kind of graph the incomplete graph should be. It may be a 3 element path, 4 element circuit, tree, k-tree, etc.

If it is desired to know only the number of graphs which are contained in each of these topologically distinct incomplete graph classes we would proceed in the following manner.

A v vertex graph could have the vertices named in $v!$ different ways if the vertices were topologically distinct. Vertices may have the same properties, that is, the same degree for the vertices and the same degree at the other end of each element connected to the vertices being compared. For the case where 2 vertices have the same topological properties, interchanging their names would not produce a different graph. The number of distinct graphs is $n!/2$. Where 3 vertices have the same properties there are 3! ways they may be renamed without producing a different graph. Therefore, the number of graphs is reduced by this factor producing $n!/3!$ graphs.

Generalizing, let n_i be the number of graphs which exist in class i. If x topologically like vertices reduce the possibilities by $1/x!$ and y topologically like vertices reduce the possibilities by $1/y!$, then

$$n_i = \frac{n!}{x!\ y!} \text{ graphs}$$

or, in general

$$n_i = \frac{n!}{\pi\ x_i\ !} \text{ graphs.}$$

ENUMERATION

Aside from the general interest in the abstract viewpoint of the subject presented thus far, an obvious use of the viewpoint is for enumeration. This will be discussed briefly in this concluding section.

Referring to Figure 3 which displayed the four topologically distinct ways to remove three elements from a 5 vertex complete graph, note the following. The dotted elements of 3a represent a 3 element circuit and 3c is a 3 element path. Neither a tree nor 2-tree is possible because of an insufficient number of elements.

In Figure 10, the compositions and graphs are shown for the 5 vertex graph with four missing elements.

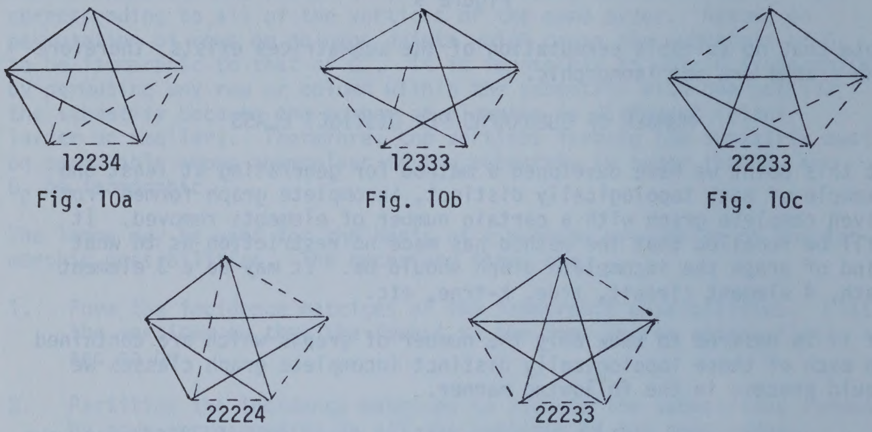

12234
Fig. 10a
12333
Fig. 10b
22233
Fig. 10c
22224
Fig. 10d
22233
Fig. 10e

Figure 10

A path, circuit, 2-tree, and tree are displayed in this set of possibilities.

The number of graphs per possibility is easily found. For example, to compute the number of three element paths in the 5 vertex graph, one would consider the ways of generating graphs isomorphic to that of Figure 3a. The composition is 222255, the number of graphs in the class is 5!/3! 2! or 10 graphs, and the transition matrix provides a means of generating each of the set.

REFERENCES

1. "Computer Determination of Isomorphisms in Linear Graphs," R.G. Goodman and R.L. Cummins, Proceedings of the 8th Midwest Symposium on Circuit Theory, June 14, 1965, pp. 12.1 - 12.12.

2. "Subgraph Identification - Segs, Circuits, and Paths," L.M. Maxwell and G.B. Reed, Proceedings of the 8th Midwest Symposium on Circuit Theory, June 14, 1965, pp. 13.1 - 13.10.

3. "Generation of Trees Without Duplication," W. Mayeda and S. Seshu, IEEE Transactions on Circuit Theory, Vol. CT-12, No. 2, June 1965.

4. "Linear Graphs and Electrical Networks," S. Seshu and M.B. Reed, Addison-Wesley Publishing Co., 1961.

5. "Intro. to Combinatorial Analysis," J. Riordan, Wiley, 1958, p.107

THE SHORTEST GENERALIZED HAMILTONIAN TREE

Narsingh Deo* and S. L. Hakimi**
*Burroughs M & E Division, Pasadena, California
**Northwestern University, Evanston, Illinois

ABSTRACT

The problem considered is that of interconnecting a given set of terminals with a shortest possible network with the added constraint that no more than a prespecified number of links be connected to any terminal. A linear programming solution of this problem is presented. It is also shown that this problem is a generalization of several long standing classical problems.

INTRODUCTION

An important problem in the backplane wiring of a large digital computer is the following:

Given a set of n points or vertices (connector pins on the backplane of the computer), connect them by a network of links (pieces of wire) having the smallest possible total length (i. e., using the smallest amount of wire), such that no more than r links are connected to any point. The last condition results from the fact that the length of each pin is limited and the wiring machine cannot wrap more than a certain number of (typically three or four) wires around one pin. The number of links (or edges) joined at a vertex is called the degree of that vertex. This problem is of inherent interest also in the planning of large scale communication, distribution, and transportation networks where the degree of each terminal is restricted due to reliability considerations [1, 2] .

When the degree constraint is removed, we have simply the problem of finding the shortest tree in a graph, and this has been solved by Kruskal [3] , and Prim [4] .

A circuit is called Hamiltonian if it passes through every vertex in the graph exactly once [5] . There have been two classically unsolved problems in this area: 1) Given an incomplete graph, how can one tell if it has a Hamiltonian circuit or not? 2) Given a complete graph, how to find the shortest Hamiltonian circuit, without having to enumerate all of $\frac{(n-1)!}{2}$ Hamiltonian circuits [5, 6] . The second problem is exactly the same as the Traveling Salesman Problem in Operations Research.

The removal of an edge from a Hamiltonian circuit produces a

Hamiltonian path. It can be shown that the removal of the longest edge from the shortest Hamiltonian circuit does not always produce the shortest Hamiltonian path in the graph, even when the lengths of edges satisfy the distance condition of a metric space*. It would be an interesting problem to find the necessary and sufficient conditions under which the relationship: Shortest H-Circuit - Its longest edge = Shortest H-Path always holds.

Since a Hamiltonian path is a tree in which the degree of no vertex exceeds two, we can define a generalized Hamiltonian tree in which the degree of no vertex exceeds a certain spacified number r; $2 \leq r \leq (n-1)$. The solution of the backplane wiring problem in a computer then is nothing but the shortest Hamiltonian tree in a graph G. G may not always be a complete graph because certain connections may be prohibited by other considerations.

PRELIMINARY NOTIONS

Let the distance between points be described by an n by n real symmetric matrix $D = (d_{ij})$, where d_{ij} represents the length of the single edge connected between points i and j. If there is no edge connected between points i and j, then $d_{ij} = \infty$. A network can be specified by an n by n symmetric "connection matrix" $X = (x_{ij})$, where $x_{ij} = 1$ whenever points i and j, $(i \neq j)$, are connected by a direct link, otherwise $x_{ij} = 0$. In the present problem, we are given matrix D, and we are required to find a matrix X such that it minimizes the total length L(x) defined as

$$L(x) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} x_{ij} d_{ij} \tag{1}$$

while satisfying conditions (2), (3), (4) and (5)

$$\sum_{k=1}^{i-1} x_{ki} + \sum_{m=i+1}^{n} x_{im} \leq r, \text{ for all i's, } 1 \leq i \leq n \tag{2}$$

$$\sum_{k=1}^{i-1} x_{ki} + \sum_{m=i+1}^{n} x_{im} \geq 1, \text{ for all i's, } 1 \leq i \leq n \tag{3}$$

---

* These conditions are

(i) $d_{ij} = d_{ji}$

(ii) $d_{ij} \leq d_{ik} + d_{ki}$

(iii) $d_{ij} = 0$, iff, $i=j$, where d_{ij} = distance of point j from i.

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} x_{ij} = (n-1) \qquad (4)$$

$$x_{ij} \leq 1 \qquad (5)$$

$$\text{the subgraph contains no loops} \qquad (6)$$

Inequality (3) implies that there is no isolated vertex in the network, and inequality (2) implies that the degree of no vertex exceeds r, where r is a given positive integer, $2 \leq r \leq (n-1)$. Equation (4) implies that the total number of links in the network should be (n-1), which is the minimum number of links required to connect n points together. To guarantee Condition (6), we have devised a separate single test for connectivity of the graph - a test which every solution must pass.

TREE TEST ALGORITHIM

A subgraph g which satisfies inequality (3) and equation (4) must have at least two pedant vertices*. If we remove all pedant vertices and the links incident on them from g, we are left with (a) a null graph, or (b) a smaller graph, with one or more pedant vertices, or (c) a smaller graph with no pedant vertex. It is easy to see that case (a) implies that the graph is a tree, and case (c) implies that the graph is not a tree. If we have case (b) we again remove all the pedant vertices and the links incident on them. Again, we obtain case (a), or (b), or (c). This process is repeated until we have either nothing left, in which case the subgraph passes the tree test, or we cannot reduce the (non zero) graph any further, in which case the subgraph fails the tree test. Furthermore, it can be easily shown that the maximum number of steps necessary in such a tree test is

$$(n-4) \text{ for } n > 9$$
$$\text{integral part of } \frac{(n+1)}{2} \text{ for } n < 9$$

In actual programming, in order to remove the pedant vertices, the computer adds the entries in each row of matrix X. The rows that have only one non zero entry are made zeros and all corresponding entries in the columns are also set to zero, because the matrix must always remain symmetric.

---

* A vertex of degree one is called a pedant vertex.

LINEAR PROGRAMMING FORMULATION
AND DIGITAL COMPUTER SCHEME

This part can be best illustrated with an example. Let us consider for simplicity a complete graph G_1 of six vertices.

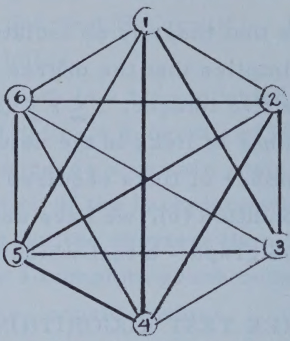


Fig. 1 - Graph G_1

The distance matrix $D = (d_{ij})$ of graph G and its connection matrix X are the following:

$$D = \begin{array}{ccccccc} & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\ v_1 & 0 & 4 & 10 & 18 & 5 & 10 \\ v_2 & 4 & 0 & 12 & 8 & 2 & 6 \\ v_3 & 10 & 12 & 0 & 4 & 18 & 16 \\ v_4 & 18 & 8 & 4 & 0 & 14 & 6 \\ v_5 & 5 & 2 & 18 & 14 & 0 & 16 \\ v_6 & 10 & 6 & 16 & 6 & 16 & 0 \end{array} \qquad X = \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\ 0 & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} \\ & 0 & x_{23} & x_{24} & x_{25} & x_{26} \\ & & 0 & x_{34} & x_{35} & x_{36} \\ & & & 0 & x_{45} & x_{46} \\ & & & & 0 & x_{56} \\ & & & & & 0 \end{array}$$

We are to find the shortest Hamiltonian tree with r = 2. The entries of the following connection matrix X are to be determined such that the objective function

$$L(x) = 4x_{12} + 10x_{13} + 18x_{14} + 5x_{15} + 6x_{16} + 12x_{23} + 8x_{24} + 2x_{25} + 6x_{26} + 4x_{34} + 18x_{35} + 16x_{36} + 14x_{45} + 6x_{46} + 16x_{56}$$

is minimum, with the degree constraints

$$x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + y_1 = 2$$
$$x_{12} + x_{23} + x_{24} + x_{25} + x_{26} + y_2 = 2$$
$$x_{13} + x_{23} + x_{34} + x_{35} + x_{36} + y_3 = 2$$
$$x_{14} + x_{24} + x_{34} + x_{45} + x_{46} + y_4 = 2$$
$$x_{15} + x_{25} + x_{35} + x_{45} + x_{56} + y_5 = 2$$
$$x_{16} + x_{26} + x_{36} + x_{46} + x_{56} + y_6 = 2$$

Since there must be no isolated vertex, $0 \leq y_i \leq 1$ for $i = 1, 2, \ldots 6$. This is done by introducing some additional variables and writing the following equations under the constraint $y_i \geq 0$ for $i = 1, 2, \ldots, 12$

$$y_i + y_{i+6} = 1, \quad i = 1, 2, \ldots, 6$$

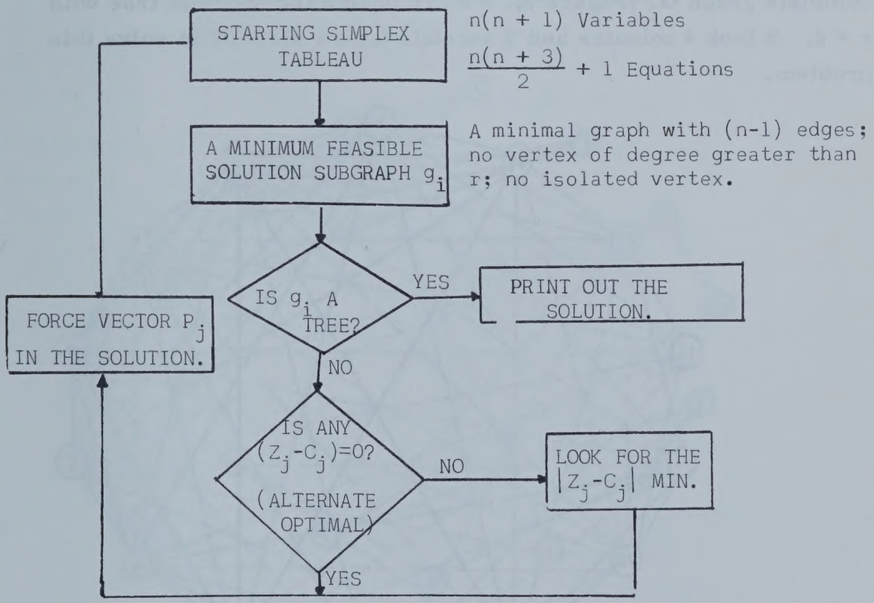
and

$$x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{23} + x_{24} + x_{25} + x_{26} + x_{34} + x_{35} + x_{36} + x_{45} + x_{46} + x_{56} = 5.$$

Also since $0 \leq x_{ij} \leq 1$, we have

$$\begin{aligned} x_{12} + Z_1 &= 1 \\ x_{13} + Z_2 &= 1 \\ \cdot \quad \cdot \quad &\cdot \\ x_{56} + Z_{15} &= 1 \end{aligned}$$

After forming a simplex tableau with $n(n+1) = 42$ variables and $\frac{n(n+3)}{2} + 1 = 28$ equations, we use the following scheme of solution.


STARTING SIMPLEX TABLEAU
n(n + 1) Variables
$\frac{n(n+3)}{2} + 1$ Equations
A MINIMUM FEASIBLE SOLUTION SUBGRAPH g_i
A minimal graph with (n-1) edges; no vertex of degree greater than r; no isolated vertex.
IS g_i A TREE?
YES
PRINT OUT THE SOLUTION.
NO
FORCE VECTOR P_j IN THE SOLUTION.
IS ANY $(Z_j - C_j) = 0$? (ALTERNATE OPTIMAL)
NO
LOOK FOR THE $|Z_j - C_j|$ MIN.
YES

Linear Programing Solution Scheme

The first solution of this problem (Figure 2 a) with a weight of 21 failed the tree test. The final solution (Figure 2b) with a weight of 23 is the shortest Hamiltonian tree.

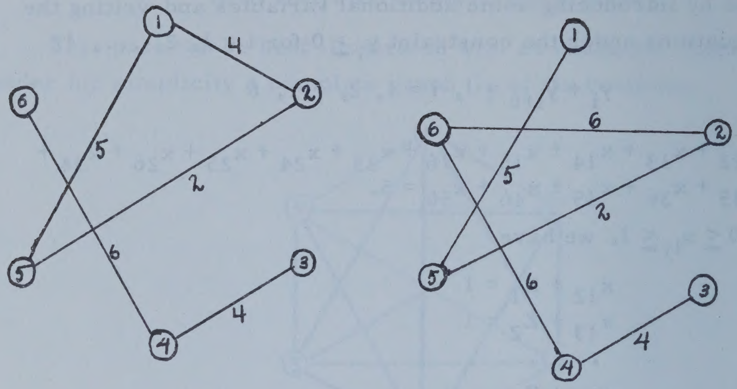

1
2
3
4
5
6
4
5
2
6
4
1
6
2
5
3
4
6
5
2
6
4

Fig. 2(a) - The Shortest Subgraph, and (b) The final Solution

After having illustrated the step by step procedure for a small graph, we now use this technique to solve a larger problem. In an incomplete graph G_2 (Figure 3), we are to find the shortest tree with r = 2. It took 4 minutes and 9 seconds for an IBM 709 to solve this problem.

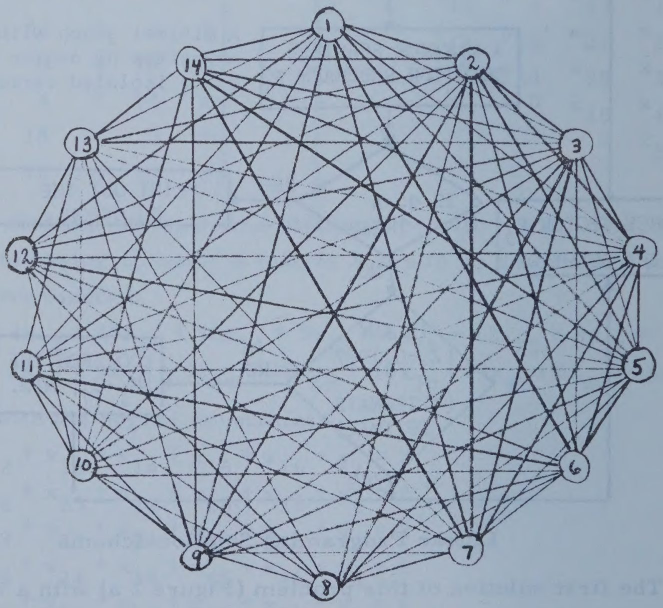

1
2
3
4
5
6
7
8
9
10
11
12
13
14

Fig. 3 - An Incomplete Linear Graph G with 14 Vertices and 82 Edges

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}
V_1	0	28	57	72	81	85	80	∞	89	80	30	38	47	17
V_2		0	28	45	54	57	63	85	63	63	34	28	15	24
V_3			0	20	30	40	57	57	40	57	35	52	18	61
V_4				0	15	20	72	45	20	45	57	49	54	53
V_5					0	22	81	41	10	50	53	50	75	82
V_6						0	63	28	28	63	55	58	∞	∞
V_7							0	80	89	∞	59	60	99	21
V_8								0	55	15	66	93	65	∞
V_9									0	40	95	∞	39	88
V_{10}										0	103	121	42	74
V_{11}											0	∞	33	∞
V_{12}												0	17	∞
V_{13}													0	70
V_{14}														0

THE D MATRIX OF GRAPH G
(∞ WEIGHT REPRESENTS A MISSING EDGE)

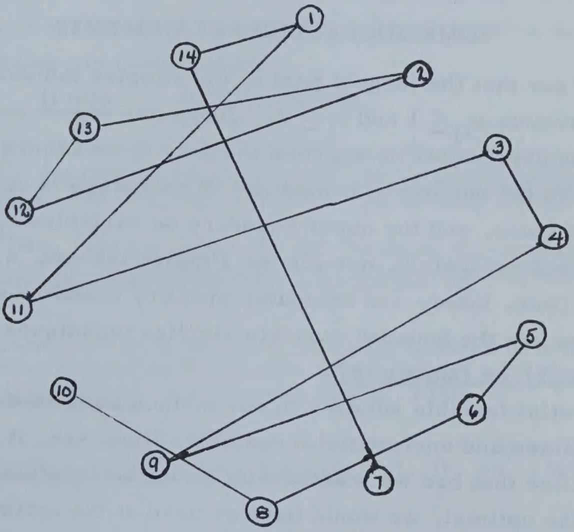


Fig. 4 - The Minimal Solution (Total Weight 278)

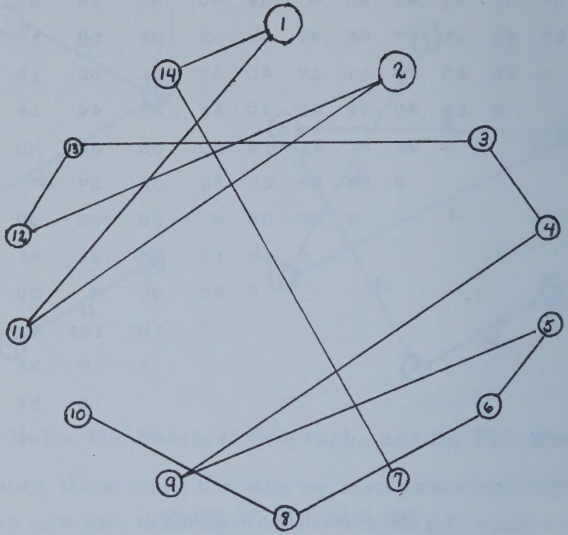


Fig. 5 - The Final Solution (Total Weight 280)

SOME SUGGESTED REFINEMENTS

It is clear that the largest part of the simplex tableau was taken up by constraints $x_{ij} \leq 1$ and $y_i \leq 1$. There are $\frac{n(n+1)}{2}$ such constraints, whereas the constraints on degrees are only n, and there is only one constraint on the number of branches. If advantage is taken of this special structure, and the upper boundary on variables x_{ij} and y_i' s were imposed separately, outside the simplex tableau, a substantial savings in time, labor, and computer memory would be realized. One can perhaps use the bounded variable simplex techniques of Charnes and Lemke [7] or Dantzig [8].

The initial feasible solution in our method consisted of $\frac{n(n+1)}{2}$ slack variables and one artificial variable. However, it is not difficult to realize that had we started with an initial feasible solution closer to the optimal, we would have arrived at the optimal solution much sooner. This initial solution can be in the form of a modified Kruskal-Prim tree, in which the vertices that do not satisfy the degree constraints are redrawn. Although this will not be the optimal solution, it will be very close. Moreover, the weight of a Kruskal-Prim tree [3, 4], gives the absolute lower bound on the weight of the solution.

The integrality of the solution has been tacitly assumed in the program, and the only condition on variables is that $x_{ij} \leq 1$. Intuitively it appears that if it is economical for a particular x_{ij} to be present in the solution, it will also be economical for it to attain its maximum allowable value (which is 1). In many respects this is similar in structure to the Transportation Problem [9], where it has been shown that solutions are always integral. We have, however, not been able to prove this rigorously for the present problem.

CONCLUSION

A problem of obtaining an optional tree with degree constraints has been proposed and solved by linear programming. Some of the well known classical problems such as Hamilton's problems, Traveling Salesman problem, Kruskal-Prim problem, are special cases of this generalized problem. The solution of this problem is of great value, not only in the backplane wiring of computers, but also in the layout of integrated circuitry, card modules, etc. An analytic solution of this problem is, indeed, a goal to strive for, because although the present linear programming method always gives a solution, it has its inherent bulkiness and inefficiency.

REFERENCES

1. Y. Fu and S. S. Yau, "A Note on Reliability of Communication Networks", Journal of the Society of Industrial and Applied Mathematics, Vol. 10, pp. 469-475, September, 1962.

2. Narsingh Deo, "On Survivability of Communication Nets", IEEE Transactions on Communication Technology, pp. 227-228, December, 1964.

3. J. B. Durskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem", Proceedings of American Mathematical Society, Vol. 7, pp. 48-50, 1956.

4. R. C. Prim, "Shortest Connection Networks and Some Generalization", Bell System Technical Journal, Vol. 26, pp. 1389-1401, November, 1957.

5. F. Supnick, "Extreme Hamilton Lines", Annals of Mathematics, Vol. 66, pp. 179-201, July, 1957.

6. O. Ore, "Note on Hamilton Circuits", American Mathematical Monthly, Vol. 67, p. 55, 1960.

7. A. Charnes and C. E. Lemke, "Computational Theory of Linear Programming: Part I, The Bounded Variable Problem", O. N. R. Research Memorandum, No. 10, Graduate School of Industrial Administration, Carnegie Institute of Technology, January 7, 1954.

8. G. B. Dantzig, "Upper Bounds, Secondary Constraints and Block Triangularities in Linear Programming", Econometrica, Vol. 23, pp. 174-183, 1955.

9. G. B. Dantzig, "Application of Simplex Method to a Transportation Problem", Activity Analysis of Production and Allocation, Edited by T. C. Koopman, J. Wiley and Sons, pp. 359-373, 1959.

GENERATION OF TREES AND k-TREES*

WAI-KAI CHEN
Department of Electrical Engineering
Ohio University
Athens, Ohio

ABSTRACT

A method, which is readily adaptable for digital computation for generating directed trees and directed k-trees of a given digraph, is presented. The method, in general, is easy to remember and simple to operate. When the digraph is symmetric, it reduces to the undirected case. Specific applications of the technique to the analysis of active electrical networks are also discussed.

1. INTRODUCTION

The efficiency with which a complex electrical network that is passive and contains no mutual inductances may be analyzed by a digital computer depends largely upon the efficiency with which the trees of the corresponding graph are generated [1-6]. When the network is active it is known that the determinants and their cofactors of the node-admittance matrix and/or the loop-impedance matrix of an active electrical network can be evaluated by finding the set of all directed trees and the set of all directed 2-trees of the corresponding directed linear graph (or simply digraph) [7]. The problem of efficient generation of all directed trees and 2-trees of a digraph is therefore important. Since an undirected graph may be transformed into a directed graph by representing each undirected edge by a pair of directed edges with opposite directions, the process of generating all trees of an undirected graph becomes a special case of the more general technique of generating all directed trees of a digraph.

This paper presents a method for generating directed trees, directed 2-trees and, in general, directed k-trees of a given digraph. The method may be considered as an extension of the idea proposed by Wang[8], Ting[9], and Wozniacki[10], and is readily adaptable for digital computation. In the last part of the paper, specific applications of the technique to various simple network structures are given.

2. DEFINITIONS AND PRELIMINARY CONSIDERATIONS

A directed linear graph G(V,E) (or simply digraph G) consists of a set V of elements called nodes together with a set E of ordered pairs of the form (i,j), i and j ε V, called the directed edges (or simply edges) of the graph; the node i is called the initial node and node j the terminal node. Together they are called the endpoints of the edge (i,j).

For any i ε V the symbols ρ(i) and ρ*(i) denote the cardinals of the sets of edges of G having i as initial and terminal nodes, respectively. They are called the outgoing and incoming degrees of G at i. A digraph is regular of degree k if ρ(i) = ρ*(i) = k for each i ε V. A digraph is symmetric if (i, j) ε E, then (j, i) ε E.

---

* The work reported in this paper was supported by the National Science Foundation under the grant number GK-173.

A graph S is a subgraph of G if the nodes and edges of S are nodes and edges respectively of G and if each edge of S has the same initial node and the same terminal node in S as in G. If the node set and the edge set of S are both empty, then S is called the empty graph ϕ of G. A node not incident to any edge is called an isolated node. A digraph consists only of isolated nodes is called the null graph. The empty graph is considered as a subgraph of every digraph. A directed circuit of G is a connected regular subgraph of degree 1. Two subgraphs are disjoint when they have no nodes in common. If S is a subgraph of G and each node of G is a node of S, then S is a spanning subgraph of G. A directed path P_{ij} is a sequence of the form $P_{ij} = (i,k_1)(k_1,k_2)(k_2,k_3) \ldots (k_m,j) \in E$ where i, j and k_t, $t = 1, \ldots, m$ are nodes in V. It is not required that all the nodes of P_{ij} shall be distinct. If they are, P_{ij} is said to be a directed simple path. Again node i is called the initial node of P_{ij} and node j the terminal node. Both nodes i and j are referred to as end nodes. A directed circuit (or a directed path) is of length m if the number of edges contained in the directed circuit (directed path) is m.

If the order of two endpoints of an edge is immaterial i.e., $(i,j) = (j,i)$, then the edge is called an undirected edge. A digraph G is said to be undirected when every edge is undirected; it is denoted by G_u. Sometimes, it is desirable to change an undirected graph G_u into a directed one G by the process of duplication: one replaces each undirected edge in G_u by two edges with the same endpoints; to the two edges in such a pair one assigns opposite directions. G is said to be connected if, for any two nodes i and j, there exists an undirected path in G_u with these two nodes as end nodes. A component of G is a maximal connected subgraph of G. Let A be a subset of V, the sectional graph G[A] of G defined by A is the subgraph whose node set is A and whose edges are all those edges in G which connect two nodes in A.

A tree t of G_u is a spanning subgraph which is connected and contains no undirected circuits. A spanning subgraph H of G_u is said to be a k-tree of G_u if H has k components, and contains no undirected circuits. Some of the components may consist of an isolated node. Very often, k-trees in which certain designated nodes are required to be in different components are used. For example, $t_{ab,cde,f}$ is the symbol for a 3-tree in which the nodes a,b are in one component, the nodes c,d,e in one component, and the node f in one component. A subgraph, denoted by t^r, of a digraph G(V,E) is said to be a directed tree if and only if (1) the corresponding undirected graph of t^r forms a tree in the corresponding undirected graph G_u of G; and (2) $\rho(i) = 1$ for all $i \in V-r$ where r is called the reference node of G and V-r is the node set with node r having been deleted from V, and $\rho(r) = 0$. A subgraph, denoted by $t^r_{r,i}$, of G is said to be a directed 2-tree if and only if (1) the corresponding undirected graph $t^r_{r,i}$ is a 2-tree in the corresponding undirected graph G_u of G; (2) each of the two components of $t^r_{r,i}$ is a directed tree for some sectional graph of G; node r serves as a reference node for one of such directed trees; and node i the other. The extension to directed k-tree is obvious. Again $t^r_{ra,jbc,k}$ is the symbol for a directed 3-tree in which the nodes r,a are in one component; the nodes j,b,c in one component; and the node k in one component. Nodes r,j,k are the reference nodes of the components.

For a given set A, $\mu(A)$ is used to denote the number of elements in A. For convenience edges of a given digraph G(V,E) will also be represented by letters e_i. Each subgraph Q of G may be expressed by the "product" or by juxtaposition of its elements provided that Q contains no isolated nodes; i.e., for example, a directed tree may be represented by $e_{i_1}e_{i_2}\dots e_{i_{\mu(V)-1}}$. Two set theoretic binary operations, "union" and "intersection" represented by "U" and "$\cap$", respectively, will be used. $g_1 U g_2$ is a subgraph of G consisting of all elements which are either in g_1 or in g_2 or in both where g_1 and g_2 are two subgraphs of G. $g_1 \cap g_2$ is a subgraph consisting of all elements which are in both g_1 and g_2. Again, if $g_1 U g_2$ contains no isolated nodes, then it may be represented by the product of its elements. Similarly for $g_1 \cap g_2$; i.e., for example, if $g_1 = e_1e_2e_3e_4$ and $g_2 = e_1e_4e_5e_6$, then $g_1 U g_2 = e_1e_2e_3e_4e_5e_6$ and $g_1 \cap g_2 = e_1e_4$.

3. PROPOSED ALGEBRA

Let D be the set of all subgraphs of a given digraph G(V,E). Consider the set H of all subsets of D for which first of all an equivalent relation "=" is defined. Assume in H there are two binary operations. One of these binary operations is called "ring-sum" and is denoted by the symbol "$\oplus$". The second of these operations is called "star" and is denoted by the symbol "*". These two operations are defined to satisfy the following rules: for $h_1, h_2 \in H$

$h_1 \oplus h_2 = \{ x \mid x \text{ in } h_1, \text{ or in } h_2, \text{ but not in both}\}$
$h_1 * h_2 = \{ xUy \mid x \in h_1;\ y \in h_2;\ \text{and } x \cap y = \phi \}$

It can easily be shown that the operations define here are commutative and associative. Furthermore, the set H forms an abelian group with respect to the "ring-sum" operation, and a commutative semi-group with unity with respect to the "star" operation. The zero element in H is the empty set of H and is denoted by $h_o = \{ \}$ while the identity element is $\{\phi\}$ where ϕ is the empty graph of G. For every $h \in H$, $h_o * h$ is defined to be h_o; and the inverse of h relative to ring-sum operation is itself, i.e., $h \oplus h = h_o$.

4. GENERATION OF DIRECTED TREES

(a) For Symmetrical Digraphs

Assume that the given digraph G(V,E) is symmetric, and all the nodes of V are labelled by one of the integers from 1 to $\mu(V)$. Let C_1 be the Cartesian product of the subsets S_k, i.e.,

$$C_1 = \prod_{\substack{k=1\\ k\neq r}}^{\mu(V)} S_k$$

where $S_k = \{ (k,t) \mid (k,t) \in E \text{ and } t = 1,2,\dots,k-1,\ k+1,\dots,\mu(V) \}$ for $k = 1,2,\dots,\mu(V)$, but $k \neq r$ where r is an arbitrary node of V. It follows that each element c of C_1 is either a directed tree t^r or a subgraph which contains at least one directed circuit of length ≥ 2. By eliminating the latter from C_1 it can easily be shown that the remaining elements of C_1 will give all possible directed trees of G. Suppose c' is an element of C_1 which contains at least one directed circuit of length > 2, then $c' = R_1 U R_2$ where $R_1 = (i_1,i_2)(i_2,i_3)\dots(i_k,i_1)$, $k > 2$, is a directed circuit of c' and R_2 is the subgraph which contains those elements of c' which have not been accounted in R_1. Since G is assumed to be symmetric, there always exists a unique $c'' \in C_1$ other than c' such that $c'' = R_3 U R_2$ where $R_3 = (i_1,i_k)\dots(i_3,i_2)(i_2,i_1)$, $k > 2$, is a directed circuit of c". It follows that the set of elements of C_1 can in general be grouped into three disjoint subsets M_t, $t = 1,2,3$, such that

$M_1 = \{ c_i \mid c_i \in C_1 \text{, and } c_i \text{ is a directed tree of G} \}$
$M_2 = \{ c_i \mid c_i \in C_1 \text{; there exists at least one directed circuit of length 2 in } c_i \}$

$M_3 = \{ c_i \mid c_i \epsilon C_1;\ c_i$ is not a directed tree of G and contains no directed circuits of length 2$\}$

Since there exists one-one correspondence between directed trees of a symmetrical digraph G and trees of the corresponding undirected graph G_u of G [7], it is simpler to consider G_u rather than G, i.e., one considers $(i,j) = (j,i)$ for all $(i,j) \epsilon E$. Consequently,

Lemma 1: For each $c \epsilon M_2$, $\{y_1\} * \ldots * \{y_{\mu(V)-1}\} = h_o\ (= \{\ \})$ where $c = y_1 y_2 \cdots y_{\mu(V)-1}$.

Lemma 2: $\{c_1\} \oplus \{c_2\} \oplus \ldots \oplus \{c_{\mu(M_3)}\} = h_o$ where $c_i \epsilon M_3$ for $i = 1, 2, \ldots, \mu(M_3)$.

Proof: Since each element of M_3 appears exactly twice in M_3, the lemma follows immediately.

Theorem 1: If G(V,E) is a symmetrical digraph, then the set T^r of all directed trees t^r is given by $T^r = \{x_1\} \oplus \{x_2\} \oplus \ldots \oplus \{x_m\}$ where $S_1 * \ldots * S_{r-1} * S_{r+1} * \ldots * S_{\mu(V)} = \{x_i,\ i = 1, 2, \ldots, m$ for some positive integer m$\}$ and $S_k = \{ (k,t) \mid (k,t) \epsilon E$ and $t = 1, 2, \ldots, k-1, k+1, \ldots, \mu(V) \}$ for $k = 1, \ldots, r-1, r+1, \ldots, \mu(V)$; and $r \epsilon V$ is the reference node of G. (i,j) and (j,i) are considered identical in the operations.

(b) For Arbitrary Digraphs

Now an arbitrary digraph G(V,E) is considered, i.e., it is no longer required that G be symmetric. In order that the method discussed in (a) be used to generate all directed trees of an arbitrary digraph G, it is necessary that G be modified first. Let G' (V',E') be the digraph obtained from G(V,E) such that $V' = V$ and $E' = \{ (i,j) \mid$ either (i,j) or $(j,i) \epsilon E\}$. It is obvious that G' (V',E') is obtained from G by adding some edges not in E. In order that one may distinguish the added edges from those in E, the following labeling process for edges in E' will be adopted: if both (i,j) and (j,i) are in E, let $(i,j) = e_k$, $(j,i) = e'_k$; if $(i,j) \epsilon E$, and $(j,i) \epsilon E'$ but not in E, let $(i,j) = e_t$ and $(j,i) = e''_t$ where k and t are some positive integers.

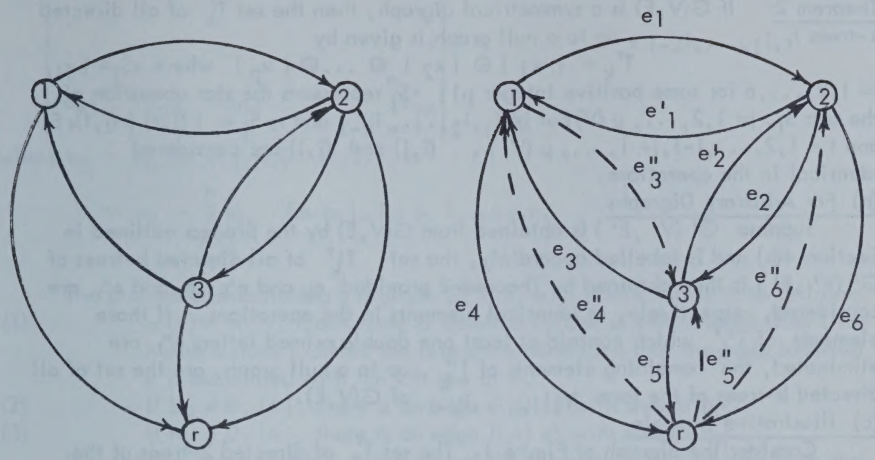
Suppose the edges of G' (V',E') are labeled according to the rules just adopted, the set T'^r of all directed trees t'^r of G' (V',E') is then obtained by Theorem 1 provided e_k and e'_k, e_t and e''_t are considered, respectively, as identical elements in the operations. It follows that if one eliminates all those elements of T'^r which contain at least one double primed letters e''_t, then the remaining elements of T'^r are the set of all directed trees T^r of G(V,E).

(c) Illustrative Example

Consider the digraph G(V,E) of Figure 1(a). The corresponding labelled G' (V',E') is shown in Figure 1(b).

$$S_1 * S_2 * S_3 = \{ e_1, e''_3, e_4 \} * \{ e'_1, e_2, e_6 \} * \{ e'_2, e_3, e_5 \}$$

$$\begin{aligned} = \{ & e_1 e_2 e_3,\ e_1 e_2 e_5,\ e_1 e_6 e'_2,\ e_1 e_6 e_3,\ e_1 e_6 e_5, \\ & e''_3 e'_1 e'_2,\ e''_3 e'_1 e_5,\ e''_3 e_2 e_5,\ e''_3 e_6 e'_2,\ e''_3 e_6 e_5 \\ & e_4 e'_1 e'_2,\ e_4 e'_1 e_3,\ e_4 e'_1 e_5,\ e_4 e_2 e_3,\ e_4 e_2 e_5, \\ & e_4 e_6 e'_2,\ e_4 e_6 e_3,\ e_4 e_6 e_5 \ \} \end{aligned}$$



(a) G(V,E) (b) G' (V' ,E')

Figure 1. A Given Digraph G(V,E) and Its G' (V' ,E')

$$T'^{r} = \{\, e_1e_2e_5\,,\ e_1e'_2e_6\,,\ e_1e_3e_6\,,\ e_1e_5e_6\,,\ e'_1e''_3e_5\,,\ e_2e''_3e_5\,,$$

$$e'_2e''_3e_6\,,\ e''_3e_5e_6\,,\ e'_1e'_2e_4\,,\ e'_1e_3e_4\,,\ e'_1e_4e_5\,,\ e_2e_3e_4$$

$$e_2e_4e_5\,,\ e'_2e_4e_6\,,\ e_3e_4e_6\,,\ e_4e_5e_6 \,\}$$

$$T^{r} = \{\, e_1e_2e_5\,,\ e_1e'_2e_6\,,\ e_1e_3e_6\,,\ e_1e_5e_6\,,\ e'_1e'_2e_4\,,\ e'_1e_3e_4\,,$$

$$e'_1e_4e_5\,,\ e_2e_3e_4\,,\ e_2e_4e_5\,,\ e'_2e_4e_6\,,\ e_3e_4e_6\,,\ e_4e_5e_6 \,\}$$

Since the edges of the form (r,j) will never appear in any directed tree, the labeling for these edges is actually unnecessary, and can be ignored in actual computation.

It should be noted that the expansion of $S_1 * S_2 * S_3$ is very simple. It amounts to forming the Cartesian product and then throwing away all those terms which have non-distinct subscripts. Furthermore, the ring-sum operation in Theorem 1 corresponds to eliminate all terms of $S_1 * S_2 * S_3$ which have the same subscripts.

5. GENERATION OF DIRECTED k-TREES

(a) For Symmetrical Digraphs

Let C_2 be the Cartesian product of the subsets S_j, i.e., $C_2 = \prod_{\substack{j=1 \\ j \neq i_1, \ldots, i_{k-1}, r}}^{\mu(V)} S_j$

where $S_j = \{\, (j,t) \mid (j,t) \in E, \text{ and } t = 1,2,\ldots,j-1,j+1,\ldots \mu(V) \,\}$ for $j = 1,2,\ldots,$ $\mu(V)$ but $j \neq i_1, i_2, \ldots, i_{k-1}$, and where r again is the reference node of G(V,E). It follows that each element c of C_2 is either a directed k-tree of the form $t_{r,i_1,\ldots,i_{k-1}}$ or a subgraph which contains at least one directed circuit of length ≥ 2. By eliminating the latter from C_2 it can easily be shown that the remaining elements of C_2 will give all possible directed k-trees of the form $t_{r,i_1,\ldots,i_{k-1}}$ of G. Similarly, one may state, Theorem 2: If G(V,E) is a symmetrical digraph, then the set T_k^r of all directed k-trees $t_{r,i_1,\dots,i_{k-1}}$, up to a null graph, is given by $T^r_k = \{x_1\} \oplus \{x_2\} \oplus \dots \oplus \{x_p\}$ where $*S_j = \{x_i, i = 1,2,\dots,p$ for some positive integer $p\}$; $*S_j$ represents the star operation of the sets S_j, $j = 1,2,\dots,\mu(V)$ but $j \neq i_1, i_2, \dots, i_{k-1}$ and r; $S_j = \{(j,t) \mid (j,t) \epsilon E$, and $t = 1,2,\dots,j-1,j+1,\dots,\mu(V)\}$. (i,j) and (j,i) are considered identical in the operations.

(b) For Arbitrary Digraphs

Suppose G' (V', E') is obtained from G(V,E) by the process outlined in Section 4(b) and is labelled accordinly, the set T'^r_k of all directed k-trees of G' (V', E') is then obtained by Theorem 2 provided e_k and e'_k, e_t and e''_t are considered, respectively, as identical elements in the operations. If those elements of T'^r_k which contain at least one double primed letters e''_t are eliminated, the remaining elements of T'^r_k, up to a null graph, are the set of all directed k-trees of the form $t_{r,i_1,\dots,i_{k-1}}$ of G(V,E).

(c) Illustrative Example

Consider the digraph of Figure 1. The set T^r_2 of directed 2-trees of the form $t^r_{r,2}$ is obtained as follows,

$$S_1 * S_3 = \{e_1, e''_3, e_4\} * \{e'_2, e_3, e_5\}$$

$$= \{e_1e'_2, e_1e_3, e_1e_5, e''_3e'_2, e''_3e_5, e_4e'_2, e_4e_3, e_4e_5\}$$

$$T'^r_2 = S_1 * S_3$$

$$T^r_2 = \{e_1e'_2, e_1e_3, e_1e_5, e_4e'_2, e_3e_4, e_4e_5\}$$

6. APPLICATIONS

(a) Topological Analysis for Active Electrical Networks

Many methods of developing topological formulae for general electrical networks are known [11-13]. However, they have undesirable features. Some have too many rules[11]; some require to find the signs for each term[12-13]. Recently, a method of evaluating general network determinants and their cofactors by means of directed trees and directed k-trees is proposed by Chen[7] which shows considerable improvement over the old ones. For completeness, important results of Chen are collected here as Theorem 3 and Theorem 4.

Let A_{nxn} be the network matrix (either the node-admittance or the loop-impedance matrix) of a given network where A is of the form

$$A = \begin{bmatrix} a_{11} & -a_{12} & -a_{13} & \cdots & -a_{1n} \\ -a_{21} & a_{22} & -a_{23} & \cdots & -a_{2n} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ -a_{n1} & -a_{n2} & -a_{n3} & \cdots & a_{nn} \end{bmatrix}$$

which, in turn, can be written in a slightly different form

$$A = \begin{bmatrix} \sum_{x=1}^{n} b_{1x} & -b_{12} & -b_{13} & \cdots & -b_{1n} \\ -b_{21} & \sum_{x=1}^{n} b_{2x} & -b_{23} & \cdots & -b_{2n} \end{bmatrix}$$

$$\begin{bmatrix} \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ -b_{n1} & -b_{n2} & -b_{n3} & \cdots & \sum_{x=1}^{n} b_{nx} \end{bmatrix}$$

where $b_{ij} = a_{ij}$ for $i \neq j$, $i,j = 1,\ldots, n$

$$= a_{ij} - \sum_{\substack{x=1 \\ x \neq i}}^{n} a_{ix} \quad \text{for } i = j,\ i,j = 1,\ldots, n.$$

The process of associating a digraph $G(V,E)$ with above matrix is as follows:

(1) $\mu(V) = n+1$. Each node is labelled by one of the integers from 1 to n plus a node r called the reference node such that the node labelled k is associated with the kth row of A.

(2) If $b_{ij} \neq 0$, $i \neq j$, there is an edge $(i,j) \in E$ with weight b_{ij}.

(3) If $b_{ij} \neq 0$, $i = j$, there is an edge $(i,r) \in E$ with weight b_{ij}.

It should be noted that b_{ii} is the sum of the elements in the ith row of A and usually has very simple form. In many cases the associated digraph may be drawn directly from the given electrical network diagram without first writing out the corresponding network matrix.

From an abstract point of view, an electrical network is a weighted digraph. For convenience, the associated digraph will be described as a 3-triple, denoted by $G(V,E,f)$, where f is a mapping function with domain in E and range in the complex field and such that $f((i,j))$ is the weight associated with the edge (i,j) of $G(V,E)$. Also it is convenient to extend f from a single edge $(i,j) \in E$ to any subgraph $R = \{x_i\}$ of $G(V,E,f)$ such that $f(R) = \prod f(x_i)$ where the mapping of an isolated node is defined to be 1 and the product is taken over all elements of R. If $Q = \{R_i\}$ is a set of subgraphs of $G(V,E,f)$ then $f(Q) = \{f(R_i)\}$.

Theorem 3: With all the notation used above, the network determinant is obtained as follows: $\det A = \sum_{t^r} f(t^r)$

where t^r is a directed tree in $G(V,E,f)$, and the summation is taken over all $t^r \in G(V,E,f)$.

Theorem 4: With all the notation used above, the ij^{th} cofactor Δ_{ij} of det A is obtained as follows: $\Delta_{ij} = \sum_{t^r_{ij,r}} f(t^r_{ij,r})$

where $t^r_{ij,r}$ is a directed 2-tree in $G(V,E,f)$, and the summation is taken over all $t^r_{ij,r} \in G(V,E,f)$.

Consider an electrical network shown in Figure 2 which represents a general potential feedback amplifier. The associated digraph $G(V,E,f)$ and $G'(V',E')$ of $G(V,E)$ are shown in Figure 3. The set of weights are given by

$$f(e_7) = f(e'_7) = y_1 ;\ f(e_1) = y_f;\ f(e'_1) = y_f - g_m ;\ f(e_2) = y_p + g_m$$

$$f(e'_2) = y_p ;\ f(e_3) = g_m ;\ f(e_4) = y_g;\ f(e_6) = y_2 ;\ f(e_5) = y_k$$

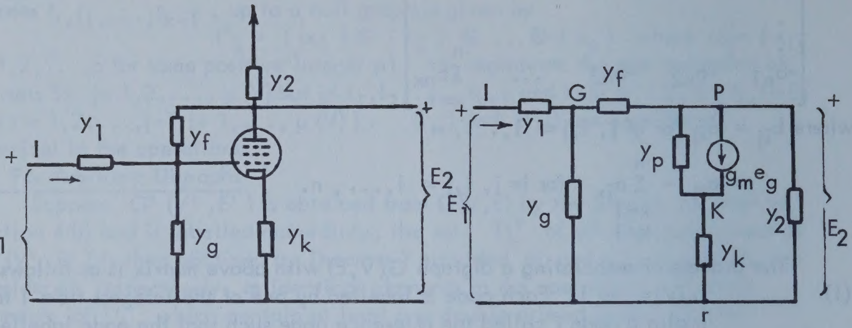
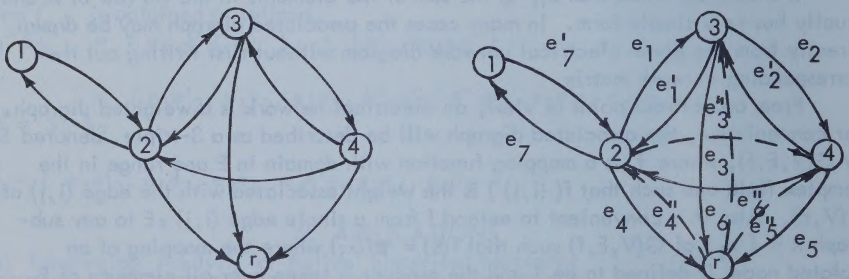


Figure 2. A General Potential Feedback Amplifier and Its Equivalent Circuit.



(a) G(V,E)

(b) G'(V',E')

Figure 3. The Corresponding G(V,E) and G' (V' ,E') of Figure 2.

$$S_1 * S_2 * S_3 * S_4 = \{e'_7\} * \{e_1, e''_3, e_4, e_7\} * \{e'_1, e_2, e_6\} * \{e'_2, e_3, e_5\}$$
$$= \{e'_7\} * \{e_1, e''_3, e_4\} * \{e'_1, e_2, e_6\} * \{e'_2, e_3, e_5\}$$

This is identical to the example considered in Section 4(c) except that e'_7 is added to each element obtained there. Therefore, the set T^r of all directed trees of G(V,E) is obtained as follows:

$$T^r = \{ e_1e_2e_5e'_7,\ e_1e'_2e_6e'_7,\ e_1e_3e_6e'_7,\ e_1e_5e_6e'_7,\ e'_1e'_2e_4e'_7,\ e'_1e_3e_4e'_7,\ e'_1e_4e_5e'_7,\ e_2e_3e_4e'_7,\ e_2e_4e_5e'_7,\ e'_2e_4e_6e'_7,\ e_3e_4e_6e'_7,\ e_4e_5e_6e'_7 \}$$

By Theorem 3, the corresponding determinant Δ of the node-admittance matrix is $\Sigma f(T^r)$ where $\Sigma f(T^r)$ represents the algebraic sum of the elements of $f(T^r)$, i.e.,

$$\Sigma f(T^r) = y_1 [y_f(y_p + g_m) y_k + y_f y_p y_2 + y_f g_m y_2 + y_f y_k y_2 + (y_f - g_m) y_p y_g$$
$$(y_f - g_m) g_m y_g + (y_f - g_m) y_g y_k + (y_p + g_m) g_m y_g + (y_p + g_m) y_g y_k +$$
$$y_p y_g y_2 + g_m y_g y_2 + y_g y_k y_2]$$

$$= y_1 [(y_k + y_p + g_m)(y_f y_2 + y_f y_g + y_g y_2) + y_k (y_f y_p + y_p y_g + y_f g_m)]$$

Similarly, the set T^r_3 of directed 2-trees of the form $t^r_{r,3}$ is obtained as follows: $T^r_3 = \{ e_1e'_2e'_7,\ e_1e_3e'_7,\ e_1e_5e'_7,\ e'_2e_4e'_7,\ e_3e_4e'_7,\ e_4e_5e'_7 \}$

By Theorem 4, the cofactor Δ_{33} is given by $\Sigma f(T^r_3)$, i.e.,

$$\Sigma f(T^r_3) = y_1 [\, y_f y_p + y_f g_m + y_f y_k + y_p y_g + g_m y_g + y_g y_k \,]$$
$$= y_1 \, (y_f + y_g)\,(y_p + g_m + y_k)$$

Hence, the output impedance Z_o looking from the points p and r is given by

$$Z_o = \frac{\Delta_{33}}{\Delta} = \frac{(y_f + y_g)\;(y_p + g_m + y_k)}{(y_k + y_p + g_m)(y_f y_2 + y_f y_g + y_g y_2) + y_k\,(y_f y_p + y_p y_g + y_f g_m)}$$

(b) Topological Analysis for Non-Recurrent Ladder Networks

Since the attempts of early pioneers in linear circuit analysis to formalize the approach to circuit problems, very little seems to have been subsequently achieved towards developing rapid methods of network analysis. The method discussed above permits a rapid analysis of the non-recurrent ladder networks. It is sufficient to illustrate these steps in terms of the ladder network of Figure 4. Suppose that the drivingpoint impedance function Z_{in} for this ladder network is required. The associated digraph G(V,E,f) and G' (V',E') of G(V,E) are shown in Figure 5.

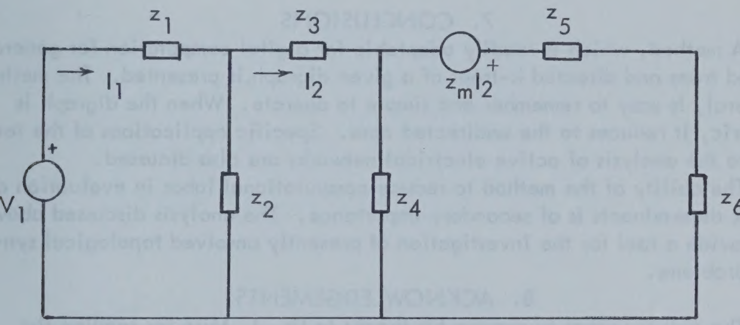
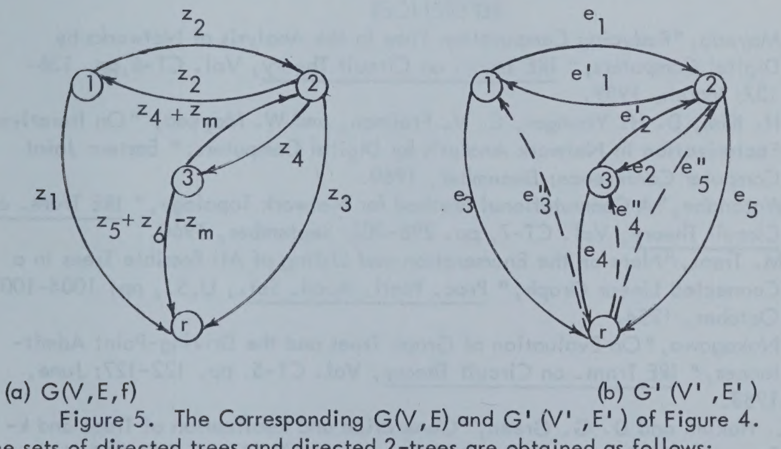


Figure 4. Ladder Network Used.



(a) G(V,E,f) (b) G' (V' ,E')

Figure 5. The Corresponding G(V,E) and G' (V' ,E') of Figure 4.

The sets of directed trees and directed 2-trees are obtained as follows:

$$S_1 {*} S_2 {*} S_3 = (e_1, e_3) * (e'_1, e_2, e_5) * (e'_2, e_4)$$
$$= (e_1 e_2,\ e_1 e_5,\ e_3 e'_1,\ e_3 e_2,\ e_3 e_5) * (e'_2, e_4)$$
$$= \{\, e_1 e'_2 e_5,\ e'_1 e'_2 e_3,\ e'_2 e_3 e_5,\ e_1 e_2 e_4,\ e_1 e_4 e_5\ ,$$

$e'_1e_3e_4,\ e_2e_3e_4,\ e_3e_4e_5\ \}$

$T^r = S_1 * S_2 * S_3$

$$\Sigma f(T^r) = (z_2z_3 + z_2z_1 + z_1z_3)(z_4 + z_m) + (z_2z_4 + z_2z_3 + z_2z_1 + z_1z_4 + z_1z_3)(z_5 + z_6 - z_m)$$

$$= (z_1z_2 + z_2z_3 + z_1z_3)(z_4 + z_5 + z_6) + z_4(z_1 + z_2)(z_5 + z_6 - z_m)$$

$$T^r_1 = \{ e'_1e'_2,\ e'_1e_4,\ e_2e_4,\ e_5e'_2,\ e_5e_4 \}$$

$$\Sigma f(T^r_1) = z_2(z_4 + z_5 + z_6) + (z_3 + z_4)(z_5 + z_6 - z_m) + z_3(z_4 + z_m)$$

$$= (z_4 + z_5 + z_6)(z_2 + z_3) + z_4(z_5 + z_6 - z_m)$$

Since the driving-point impedance function Z_{in} is defined as $Z_{in} = \frac{V_1}{I_1}$ which is equal to $Z_{in} = \frac{\Delta_{11}}{\Delta} = \frac{\Sigma f(T^r_1)}{\Sigma f(T^r)}$

The dependent voltage source z_mI_2 in Figure 4 may be part of a transistor circuit. Obviously the technique used here can be extended to more general cases.

7. CONCLUSIONS

A method, which is readily adaptable for digital computation for generating directed trees and directed k-trees of a given digraph, is presented. The method, in general, is easy to remember and simple to operate. When the digraph is symmetric, it reduces to the undirected case. Specific applications of the technique to the analysis of active electrical networks are also dicussed.

The ability of the method to reduce computational labor in evaluation of network determinants is of secondary importance. The analysis discussed above may provide a tool for the investigation of presently unsolved topological synthesis problems.

8. ACKNOWLEDGEMENTS

The author wishes to express his thanks to Dr. J. Miro for reading the manuscript and making many useful suggestions. He would also like to thank the National Science Foundation for the support of this research.

REFERENCES

1. W. Mayeda, "Reducing Computation Time in the Analysis of Networks by Digital Computers," IRE Trans. on Circuit Theory, Vol. CT-6, pp. 136-137; March, 1959.
2. W. H. Kim, D. H. Younger, C. V. Freiman, and W. Mayeda, "On Iterative Factorization in Network Analysis by Digital Computers," Eastern Joint Computer Conference; December, 1960.
3. H. Watanabe, "A Computational Method for Network Topology," IRE Trans. on Circuit Theory, Vol. CT-7, pp. 296-302; September, 1960.
4. H. M. Trent, "Note on the Enumeration and Listing of All Possible Trees in a Connected Linear Graph," Proc. Natl. Acad. Sci., U.S., pp. 1004-1007; October, 1954.
5. N. Nakagawa, "On Evaluation of Graph Trees and the Driving-Point Admittances," IRE Trans. on Circuit Theory, Vol. CT-5, pp. 122-127; June, 1958.
6. S. L. Hakimi and D. G. Green, "Generation and Realization of Trees and k-Trees," IEEE Trans. on Circuit Theory, Vol. CT-11, pp. 247-255; June, 1964.
7. W. K. Chen, "Topological Analysis for Active Networks," IEEE Trans. on Circuit Theory, Vol. CT-12, pp. 85-91; March, 1965.

8. K. T. Wang, "On a New Method for the Analysis of Networks," National Research Institute of Engineering, Academia Sinica, Memoir No. 2, pp. 1-11; 1934.
9. S. L. Ting, "On the General Properties of Electrical Network Determinants," Chinese Journal of Physics, Vol. 1, pp. 18-40; 1935.
10. H. Wozniacki, "Obliczanie i analiza sieci elektrycznyeh metoda wielomianow charakterystycznych," Archiwum Elektrotechn, Tom X, zeszyt 1, pp. 57-99; 1961.
11. S. J. Mason, "Topological Analysis of Linear Non-reciprocal Networks," Proc. IRE, Vol. 45, pp. 829-838; June, 1957.
12. C. L. Coates, "General Topological Formulas for Linear Network Functions," IRE Trans. on Circuit Theory, CT-5, pp. 30-42; March, 1958.
13. W. Mayeda, "Topological Formulas for Active Networks," Int. Tech. Rept., No. 8, U.S. Army Contract No. DA-11-022-ORD-1983, University of Illinois, Urbana, Illinois; January, 1958.

THE NUMBER OF TREES IN A NETWORK

P. V. O'NEIL and P. SLEPIAN
University of Minnesota
Rensselaer Polytechnic Inst. and Sandia Corporation

1. Summary

In this paper we exhibit an explicit formula for the number of trees in a certain type of network. In particular, we define the concept of an incompletely partitioned network, and offer as our main theorem a formula for the number of trees of an incompletely partitioned network. Our formulas include, as special cases, the formulas exhibited previously in the literature by Weinberg [1] and Bedrosian [2] .

Our main result depends upon a complicated sequence of lemmas and theorems, each of which involves the determinant of a special square matrix. Because of space limitations, we cannot exhibit these preliminary lemmas and theorems here. In this paper we shall merely define our terms and state our main result without proof. A complete derivation of our main result will appear elsewhere [3] .

2. Complete Vertices

A finite network in its most abstract form can be considered as an ordered pair of two finite sets (V,B) with V the set of vertices and B the set of branches. Each branch, $b \in B$, is itself a set consisting of exactly two vertices of V, say $x \in V$ and $y \in V$, with $x \neq y$. Thus, $b = \{x\} \cup \{y\}$. In this case we say that the branch b is incident upon the vertex x and the branch b is incident upon the vertex x and the branch b is incident upon the vertex y. We also say that the branch b connects the vertices x and y.

Suppose now that (V,B) is a network and let $v \in V$. We say that the vertex is a complete vertex if and only if it is connected by a branch to every other vertex of the network. Thus, if V consists of m vertices, the vertex v is complete if and only if exactly m-1 branches are incident upon v.

If every vertex of the network is a complete vertex, we say that the network is a complete network. Cayley [4] has shown, long ago, that the number of trees in a network with m vertices, with $m \geqslant 2$, is equal to m^{m-2}.

Much of the modern effort in tree counting has as its starting point Cayley's formula for the number of trees in a complete network. A few branches are removed, rendering the network slightly incomplete, and in certain cases the number of trees of the resulting incomplete network can be specified. Weinberg [1] and Bedrosian [2] have developed formulas for the number of trees in a network which is slightly incomplete.

Our approach will be similar, but we shall allow the removal of a substantial number of branches from the complete network. The results of Weinberg [1] and Bedrosian [2] will be special cases of our more general results.

3. Incompletely Partitioned Networks

Consider an incomplete network (V,B). Let n be a positive integer and suppose that the set of incomplete vertices of V can be partitioned into 2n subsets, A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, such that the 2n subsets are pairwise disjoint. Suppose further that there is no branch $\{x\} \cup \{y\} \in B$ such that for some $i = 1,2,\ldots n$, $x \in A_i$ and $y \in D_i$. A network (V,B) lending itself to such a partition of the set of incomplete vertices will be called an incompletely partitioned network.

Since we are originating this term we exhibit a formal definition below.

Definition 1

Let (V,B) be a network.
Then, we say that (V,B) is an incompletely partitioned network if and only if there exists a positive integer n and 2n sets A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, such that

(1) $A_i \subset V$ and $D_i \subset V$ for $i = 1,2,\ldots n$

(2) The 2n sets A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, are non empty and pairwise disjoint.

(3) $\{x\} \cup \{y\} \in B$

if and only if

there is no integer i, with $i = 1,2,\ldots n$ such that

$x \in A_i$ and $y \in Di$.

In the above situation we say that the 2n sets A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, form an incomplete partition of V, the set of vertices.

Suppose now that (V,B) is an incompletely partitioned network and we have 2n sets A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, forming an incomplete partition of V. Observe that there is noting cited in Definition 1 justifying the use of the word "incompletely". We point out trivally in Theorem 1 that the set of all complete vertices of V is equal to $V - \bigcup_{i=1}^{n}(A_i \cup D_i)$. This fact is equivalent to the statement that a vertex $x \in V$ is not a complete vertix if and only if for some $i = 1,2,\ldots n$, $x \in A_i$ or $x \in D_i$.

Theorem 1

Let (V,B) be an incompletely partitioned network and let the 2n sets A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, form an incomplete partition of V.
Then,
the set of all complete vertices of V is equal to

$$V - \bigcup_{i=1}^{n} (A_i \cup D_i).$$

As our main result we now exhibit as Theorem 2 a formula for the number of trees in an incompletely partitioned connected network.

Theorem 2

Let (V,B) be an incompletely partitioned, connected network with m vertices.
Let the 2n sets, A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, form an incomplete partition of V.
Let b_i, $i = 1,2,\ldots n$, and c_i, $i = 1,2,\ldots n$, be positive integers such that

A_i has b_i elements, if $i = 1,2,\ldots n$.
D_i has b_i elements, if $i = 1,2,\ldots n$.

$$\text{Let } k = \sum_{i=1}^{n} (b_i+c_i).$$

Then,
the number of trees in (V,B) is equal to

$$m^{m+n-k-2} \prod_{j=1}^{n} \left((m-b_j)^{c_j-1} (m-c_j)^{b_j-1} (m-b_j-c_j) \right) .$$

4. The Identification Problem

If a connected network (V,B) is assigned, our formula is not of much use unless we can determine rather easily whether (V,B) is an incompletely partitioned network. A study of the definition discloses that this property appears to be more intimately associated with the branches lacking in a network, rather than with the branches present.

This is indeed the case. Any incompletely partitioned network can be manufactured from a complete network by selecting certain disjoint sets of vertices, dividing these sets of certices into pairs, and then removing all those branches connecting a vertex of one member of a pair with a vertex of the other member of the same pair. Conversely, any such manufacturing process applied to a complete network yields an incompletely partitioned network.

We state these vague facts more precisely as two theorems. Each theorem is proved trivially by referring to Definition 1.

Theorem 3

Let (V,B) be an incompletely partitioned network. Let the 2n sets A_i, i = 1,2,...n, and D_i, i = 1,2,...n, form an incomplete partition of V.
Let (V,G) be a complete network.

Then

$$B = G - \left\{ \{x\} \cup \{y\} \,\middle|\, \text{for some } i,\ i = 1,2,\ldots n, \text{ and } \begin{array}{l} x \in A_i \text{ and} \\ y \in D_i \end{array} \right\}.$$

Theorem 4

Let (V,G) be a complete network.
Let A_i, i = 1,2,...n, and D_i, i = 1,2,...n, be 2n sets such that

(1) $A_i \subset V$ and $D_i \subset V$ if i = 1,2,...n.

(2) The 2n sets A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, are non-empty and pairwise disjoint.

Let

$$B = G - \{(x)\cdot(y \mid \text{for some } i,\ i = 1,2,\ldots n \text{ and } x \in A_i \text{ and } y \in D_i\}.$$

Then

(V,B) is an incompletely partitioned network and the 2n sets A_i, $i = 1,2,\ldots n$, and D_i, $i = 1,2,\ldots n$, form an incomplete partition of V.

Theorem 4 may be of some use in the solution of the identification problem. If a connected network (V,B) is assigned, and if the complete network (V,B) is constructed, it may be apparent that (V,B) could be manufactured from (V,G) by a process specified in Theorem 4, guaranteeing that (V,B) is an incompletely partitioned network. However, in all honesty, we must admit that this salubrious situation will not occur very often. In a complicated situation it may be very difficult to determine whether (V,B) could be manufactured from (V,G) by a process specified in Theorem 4.

Thus, a more powerful identification test is needed to determine whether an arbitrary connected network is an incompletely partitioned network. The writers have considered this problem and have produced necessary and sufficient conditions for a network to be incompletely partitioned. Furthermore, these conditions are rather wasy to apply so that an arbitrary connected network can be tested efficiently to determine whether it is incompletely partitioned. These facts, exhibiting a solution of the identification problem, are discussed and proved in detail in another paper [5].

References Cited

[1]	Weinberg, L.	"Number of Trees in a Graph," Proc. I.R.E., Vol. 46, No. 12, Dec. 1958.
[2]	Bedrosian, S.	"Formulas for the Number of Trees in a Network," I.R.E. Transactions on Circuit Theory, Vol. PGCT, No. 3, Sept. 1961.
[3]	O'Neil, P. V., and Slepian, P.	"The Number of Trees in a Network," submitted to transactions of the I.E.E.E. Professional Group on Circuit Theory.
[4]	Cayley, A.	"A Theorem on Trees," Quart. Math. Journ 1889
[5]	O'Neil, P. V. and Slepian, P.	"The Identification of an Incompletely Partitional Network," Submitted to S.I.A.M. Journal.

THE NUMBER AND LISTING OF ALL TREES IN AN ARBITRARY GRAPH

B. R. MYERS & L. V. AUTH, JR.
University of Notre Dame
Notre Dame, Indiana

I. INTRODUCTION

Product of Vertex Cutsets

The fundamental technique in Marian Piekarski's method of listing all the trees of an arbitrary graph, given in recent correspondence in the IEEE Transactions on Circuit Theory,[1] is identical with that published previously by St. Bellert.[2] Piekarski developed an astute mechanization, however, for machine computation, which circumvents the well known, time-consuming problem of searching for duplications and other unwanted terms which inevitably arise in tree generation and identification.

The basis of the technique is to write down all of the terms in the element-products expansion of any (n-1) vertex cutsets of the given n-vertex, connected graph, that have exactly one element from each cutset. On eliminating all of those terms which appear with even duplicity, and also all of those involving squared elements (this being the highest power with which any element can appear in any product - as will become apparent later), the remaining set of terms uniquely and unambiguously identifies the set of all trees of the graph.

An Example

A simple example to illustrate the mechanics of the method is given by computing all the trees of the 4-vertex graph of Figure 1, thus:

$$\begin{array}{ll} \text{Vertex} & \text{Cutset} \\ v_1 & P_1 = (a_1a_2) \\ v_2 & P_2 = (a_1a_2a_3) \\ v_3 & P_3 = (a_3a_4) \\ v_4 & P_4 = (a_4) \end{array} \qquad (1)$$

Suppose the first three cutsets are selected, with the product

$$P_1P_2P_3 = (a_1a_2)(a_1a_2a_3)(a_3a_4) \qquad (2)$$

The twelve terms in the product, with the terms to be deleted shown with stroke lines are

$$\begin{array}{llll} \cancel{a}_1^2\cancel{a}_3 & \cancel{a}_1^2\cancel{a}_4 & \cancel{a}_1\cancel{a}_2\cancel{a}_3 & \cancel{a}_1\cancel{a}_2\cancel{a}_4 \\ \cancel{a}_1\cancel{a}_3^2 & a_1a_3a_4 & \cancel{a}_2\cancel{a}_1\cancel{a}_3 & \cancel{a}_2\cancel{a}_1\cancel{a}_4 \\ \cancel{a}_2^2\cancel{a}_3 & \cancel{a}_2^2\cancel{a}_4 & \cancel{a}_2\cancel{a}_3^2 & a_2a_3a_4 \end{array} \qquad (3)$$

The remaining terms, $a_1a_3a_4$ and $a_2a_3a_4$, identify the trees of the graph. Of course, it would have been more efficient to have selected the product $P_1P_3P_4 = (a_1a_2)(a_3a_4)(a_4)$, since it yields only four terms on on expansion and gives the same end result.

Relation to the Determinant of the Node-Admittance Matrix.

This simple example also suffices to illustrate that the method is simply another way of evaluating the determinant Δ_n of the node-admittance matrix Y_n of a pssive network N which does not contain mutual inductances; for it is well known that[3]

$$\Delta_n = \sum_{\substack{\text{all} \\ \text{trees}}} (\text{tree-admittance products of N}) \tag{4}$$

The node-admittance matrix corresponding to the graph of Figure 1 is

$$\begin{bmatrix} (a_1+a_2) & -(a_1+a_2) & 0 \\ -(a_1+a_2) & (a_1+a_2+a_3) & -a_3 \\ 0 & -a_3 & (a_3+a_4) \end{bmatrix} \tag{5}$$

The entries on the main diagonal are the vertex cutsets P_1, P_2, P_3. Evaluation of the determinant of the matrix involves writing down the twelve terms listed earlier that appear in the cutset product $P_1P_2P_3$, with the deletions in this product being exactly those terms subtracted in the determinant due to the symmetric entries off the main diagonal. It is not difficult to show that this holds for a node-admittance matrix of any order.

II. ENUMERATION

As stated earlier, Piekarski's computational procedure generates all of the trees without duplication or redundant terms. The squared-element and even multiplicity terms are automatically eliminated as the method proceeds. His technique is thus particularly appropos to graphs in which structural symmetry or repetition are either not present or cannot be identified. It does not, however, give any means for knowing the total number of trees, except by counting them after they have been generated. This is presumably not difficult to do from a machine printout. In the present paper, nevertheless, it is next shown how the vertex cutset product facilitates the derivation of closed formulas for the total number of distinct trees in certain symmetrical or repetitive graphs. Specifically, the following didactic cases are considered: (1) complete graphs, (2) the subgraphs which result on removing a complete graph from another complete graph with a larger number of vertices, and (3) cascaded identical two-ports.

Complete Graphs

A labelling scheme for the elements is chosen as depicted in Figure 2, where vertex v_a is chosen as the starting point for the labelling, with its incident elements a_1, a_2, ... The vertex to which element a_1 is incident is labelled v_b, and the remaining elements incident at v_b labelled b_1, b_2, ... And so on. Whence, the vertex cutset product has a systematic structure, thus:

$$P_1P_2P_3 \dots = (a_1a_2a_3 \dots)(a_1b_1b_2 \dots)(a_2b_1c_1 \dots) \dots \dots \tag{6}$$

There are (n-1) factors in an n-vertex graph, and each factor contains (n-1) elements. Each element appears in exactly two different factors. The kth cutset P_k has the form

$$P_k = (a_{k-1}b_{k-2}c_{k-3} \ldots k_1k_2k_3 \ldots) \tag{7}$$

The other appearance of each of the first (k-1) elements a_{k-1}, b_{k-2}, c_{k-3}, ... , is respectively in, and only in, the preceeding factors P_1, P_2, ..., P_{k-1}. And the other appearance of each of the remaining elements k_1, k_2, ..., terminating at k_{n-k}, is respectively, and only in, the succeeding factors P_{k+1}, P_{k+2}, ...

It is both convenient and of applicable interest to first compute the number of distinct terms in the product N_k of the first k terms which are left after eliminating squared-element terms and terms which appear with even multiplicity. The previous observations make it easy to see that this is given by

$$\begin{aligned} N_k &= (P_1P_2 \ldots P_k) \text{ less squared terms and even duplications} \\ &= N_{k-1}N_1 - \binom{k-1}{1}N_{k-2} - 2\binom{k-1}{2}(N_1+1)^{k-3} \end{aligned} \tag{8}$$

where $\binom{x}{y}$ denotes the number of combinations of x distinct objects taken y at a time, and $N_0=1$ is assumed.

The first of the three terms on the right-hand side of (8), $N_{k-1}N_1$, is the total number of product terms due to insertion of the kth factor, assuming that N_{k-1} is known. The second term $\binom{k-1}{1}N_{k-2}$ is the number of squared-element terms which have to be deducted from $N_{k-1}N_1$ due to the first (k-1) elements in P_k, which, as described earlier, each appear once and only once in each of the respective, preceding (k-1) factors. The third term $2\binom{k-1}{2}(N_1+1)^{k-3}$ similarly accounts for the dual appearances due to each pair of these (k-1) elements, there being two eliminations for each of the $\binom{k-1}{2}$ combinations of them.

Application of the theory of recurrence sequences[4] enables the equation for N_k to be written in closed form, after some algebraic manipulation (it being noted that $N_1 = (n-1)$), as

$$N_k = (n-k)n^{(k-1)} \tag{9}$$

Putting k = (n-1) thus gives the number of distinct trees in a complete graph as

$$N_{n-1} = n^{n-2} \tag{10}$$

which is Cayley's well-known formula.

Subgraphs Resulting on Removal of a Complete Graph from a Complete Graph with a Larger Number of Vertices.

Without any loss of generality, it may be assumed that a complete graph with (n-k) vertices is removed from the complete graph with n vertices; the reduced vertices in the new graph being, in terms of the labelling scheme described earlier for Figure 2, vertices v_{k+1}, v_{k+2}, ... , v_n. The new vertex cutset product is then

$$P_1P_2 \ldots P_kP'_{k+1}P'_{k+2} \ldots P'_{n-1} \tag{11}$$

where each of the factors $P_1, P_2, \ldots, P_k$ is exactly the same as for the original n-vertex complete graph, and the primed factors each comprise only the first (k-1) elements of the corresponding factors in the original graph. For example, the new vertex cutset product of the graph obtained on removing a 3-vertex complete graph from a 6-vertex complete graph is

$$(a_1a_2a_3a_4a_5)(a_1b_1b_2b_3b_4)(a_2b_1c_1c_2c_3)(a_3b_2c_1)(a_4b_3c_2)$$

Denoting the number of trees in the reduced graph by N_k', it is a fairly straightforward matter, using the technique of the previous section, to show that

$$N_k' = N_k - k^{n-k-1}n^{k-1} \tag{12}$$

where, as before, N_k is the number of trees in the original graph. From which it can be written

$$N_k' = k^{n-k-1}n^{k-1} \tag{13}$$

This formula checks with one reported previously by Bedrosian.[5]

A Chain of Identical Two-Ports

Recurrence formulas for the number of trees in a chain of identical two-port have been given very recently by W. Gruszczynski.[6] His formulas, derived by a radically different technique, can be confirmed by the development now to be given using vertex cutsets. Further, the method here gives a series formulation in the first instance, not immediately available by Gruszczynski's method. The series formulation has an advantage in that it makes for a very clear understanding of the topological structure of the chain.

It suffices to first develop the formula for a simple example, and then show, by induction, that it holds for the general case. Figure 3 depicts a chain of identical, simple three-terminal T-network graphs. They are labelled a, b, c, ..., in that order for clarity in identification, and their respective elements are correspondingly labelled a_1, a_2, a_3, b_1, b_2, b_3, as shown. The vertex cutset product P_a to be used for network a is

$$P_a = (a_1)(a_1a_2a_3)(a_2) \tag{14}$$

For convenience, this will be written

$$P_a = A_1A_2A_3 \tag{15}$$

where

$$\begin{aligned} A_1 &= (a_1) \\ A_2 &= (a_1a_2a_3) \\ A_3 &= (a_2) \end{aligned} \tag{16}$$

Correspondingly, the vertex cutsets chosen for networks b, c, ... , are

$$P_b = (b_1)(b_1b_2b_3)(b_2) = B_1B_2B_3$$

$$P_c = (c_1)(c_1c_2c_3)(c_2) = C_1C_2C_3 \qquad (17)$$

and so on.

For later use, note that the product A_1A_2, after the squared terms and even multiplicities are deleted, gives the number of 2-tree $W_{(a)2,2'}$ of network a with respect to the right-hand terminal pair (2,2') shown in the Figure), and the product A_2A_3 gives the number of 2-trees $W_{(a)1,1'}$, (1,1') being the left-hand terminal pair. The number of 2-trees of a given kind being the same for each network, for later convenience we will write

$$W_{(a)1,1'} = X_1 \qquad (= W_{(b)1,1'} = W_{(c)1,1'} = \dots) \qquad (18)$$

and

$$W_{(a)2,2'} = Y_1 \qquad (= W_{(b)2,2'} = W_{(c)2,2'} = \dots) \qquad (19)$$

That is to say, X_1 and Y_1 will be used to denote the number of 2-trees (1,1') and 2-trees (2,2'), respectively, in each network. Further, the symbol Z_1 will be used to denote the number of 3-trees (1,1'2',2) plus (1,1'2,2') plus (12',1',2) plus (12,1',2') in each network. Quantity Z_1 is given by

$$Z_1 = A_1 + A_2 + A_4 \qquad (20)$$

where $A_4 = (a_2)$ is the one remaining cutset in network a, not considered in P_a, and where the expression for Z_1 is to be interpreted as the number of terms appearing on the right-hand side after deletion of the squared-element and even multiplicity terms, in the general case.

The vertex cutset product $P_{(n)}$ (not to be confused with the symbol P_n for the nth network) for a chain of n networks is given by

$$P_{(n)} = A_1A_2(A_3B_1)B_2(B_3C_1)C_2(C_3D_1)D_2(D_3E_1) \dots \quad \dots \qquad (21)$$

Evaluation of the product, with the requisite eliminations, is a matter of evaluating the 2^{n-1} binary products depicted by the binary tree shown in Figure 4. Careful bookkeeping reveals that N_n, the number of trees in the n-chain, is given by

$$\begin{aligned} N_n = N_{n-1}X_1 &+ N_{n-2}Z_1N_1 + N_{n-3}Z_1Y_1N_1 + N_{n-4}Z_1Y_1^2N_1 \\ &+ \dots \qquad (22) \\ &+ N_2Z_1Y_1^{n-4}N_1 + N_1Z_1Y^{n-3}N_1 + Y_1^{n-1}N_1 \end{aligned}$$

That is

$$N_n = N_{n-1}X_1 + \sum_{j=1}^{n-1} N_jZ_1Y_1^{n-j-2}N_1 + Y_1^{n-1}N_1 \qquad (23)$$

Very little algebraic manipulation leads to the recurrence formula given by Gruszczynski, viz,

$$N_n = (X_1 + Y_1)N_{n-1} + (Z_1N_1-X_1Y_1)N_{n-2}$$

where $N_o = 0$ is assumed. (24)

Gruszczysnki postulated this equation as the starting point in proving, by induction, that it holds for an n-chain of identical, but otherwise arbitrary, two-ports, whether 3-terminal or 4-terminal. Exactly the same equation holds, with exactly the same coefficients, when N_n is regarded as the number of 2-trees(1,1') or the 2-trees (2,2') of the n-chain; except that the boundary condition in these two cases is $N_o = 1$ instead of $N_o = 0$.

III. BIBLIOGRAPHY

1. Marian Piekarski, "Listing of all possible trees of a linear graph," IEEE Trans. on Circuit Theory, v. Ct-12, March, 1965, pp. 124-5.

2. St. Bellert, "Topological considerations and synthesis of linear networks by means of the method of structural numbers," Archiqum Elektrotechniki, z. 3, t. XII, 1963, pp. 473-500.

3. S. Seshu & M. B. Reed, "Linear graphs and electrical networks," Addison-Wesley Publishing Co., 1961.

4. J. Naas & H. L. Schmid, "Mathematisches Vorterbuch," Band II, Akademie-Verlag Gmbh, Berlin, 1961.

5. S. D. Bedrosian, "Generating formulas for the number of trees in a graph," Jl. of the Franklin Inst., v. 277, April, 1964, pp. 313-326.

6. W. Gruszczynski, "Number of trees in a chain of identical graphs," to appear in IEEE Trans. on Circuit Theory.

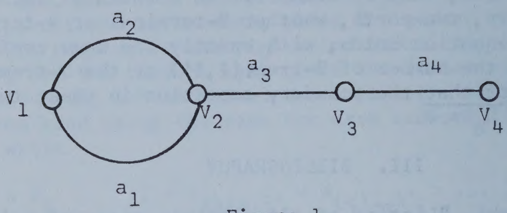

a2
a3
a4
V1
V2
V3
V4
a1

Figure 1.

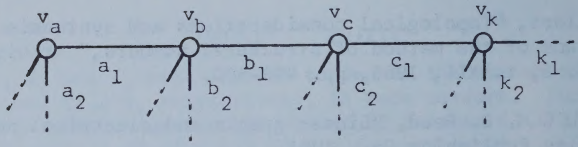

va
vb
vc
vk
a1
b1
c1
k1
a2
b2
c2
k2

Figure 2.

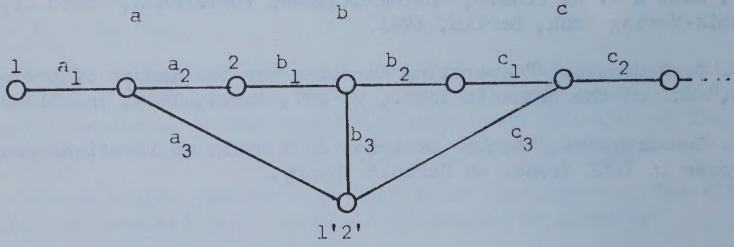

a
b
c
1
a1
a2
2
b1
b2
c1
c2
a3
b3
c3
1'2'

Figure 3.

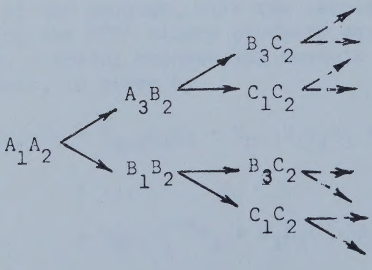

B3C2
A3B2
C1C2
A1A2
B1B2
B3C2
C1C2

Figure 4.

ALGORITHMS FOR DETERMINING AUTOMORPHISM GROUPS FOR PLANAR GRAPHS[†]

LOUIS WEINBERG
Department of Electrical Engineering
City College of the City University of New York
New York, N. Y.

ABSTRACT

Algorithms are given for determining the automorphism group of a triply conected graph, a tree, and a series-parallel graph. These algorithms are based on codes previously formulated by the author. The group of a triply connected graph is the set of permutations corresponding to a set of equal codes in the code matrix. The group of a tree is made evident by the equal consecutive subcodes operated on by an operator in a Polish-notation code. The group of a series-parallel graph is also made evident by equal consecutive subcodes, but we must properly interpret the two different operators in the Polish-notation code. Though the discussion is given in terms of undirected connected graphs, it is indicated how the procedures may be applied to unconnected graphs and to directed graphs.

INTRODUCTION

We assume the reader has a knowledge of the preceding paper on codes for planar graphs[1]; the algorithms in this paper are based on those codes. First we present a few additional definitions pertinent to the subject of this paper and give some discussion of the codes to supplement that in the previous paper. Again because of space limitations proofs will not be given here but can be found in the references and future papers to be published.

In this paper we are concerned with the problem of determining the automorphism group of a planar graph. We therefore define the term below.

Two graphs G_1 and G_2 are said to be isomorphic if there is a one-to-one correspondence between the nodes of G_1 and G_2 which preserves the incidence relations; thus in isomorphic graphs G_1 and G_2, if a branch exists between a pair of nodes a_1 and b_1 in G_1, a corresponding branch exists between the corresponding pair of nodes a_2 and b_2 in G_2, and conversely. For directed graphs the directions of corresponding branches must also be the same. Corresponding nodes of isomorphic graphs are called similar nodes; corresponding branches, circuits, and subgraphs are called similar branches, circuits, and subgraphs, respectively. An automorphism of a graph G is an isomorphism of G with itself, and the automorphism group (or briefly, the group) of G, denoted by $\Gamma(G)$ or Γ, is the collection of all its automorphisms.

The group Γ is a permutation group acting on the set $N = \{1, 2, \ldots, n\}$, where n is the number of nodes of G. The degree and order of Γ, denoted respectively by $|N|$ and $|\Gamma|$, are the cardinalities of N and Γ, respectively.

[†] This research was jointly supported by the Radar Laboratory under Project Michigan (Department of the Army, Contract No. DA 28-043-AMC-00013(E)) and by the Institute of Science and Technology, The University of Michigan.

The problem of determining the group of a graph may be stated in matrix terms. It is required that we find the group of permutation matrices P that are congruent automorphs of the node-node incidence matrix of the graph. By a congruent automorph is meant the permutation matrix P that satisfies

$$P'CP = C \tag{1}$$

where C is the node-node incidence matrix and P' is the transpose of P. In other words, P is the set of permutation matrices that commute with C. Put in these terms we see that a simple algorithm is already available; though simple to state, however, it presents insurmountable combinatorial difficulties. We need merely arbitrarily label the n nodes with the numbers 1, 2, ..., n and then determine C. We then permute this labelling, determine the new incidence matrix, and check whether it is equal to C. However, there are n! possible permutations so that, as Unger[2] points out, using a very fast computer that requires one millisecond for each trial (where a trial includes the permutation of node labelling, the formulation of the corresponding incidence matrix, and checking the two matrices for equality), the complete set of trials required to find the automorphism group for a 15-node graph would take the computer 40 years.

Though we consider only connected undirected planar graphs, our results are applicable to planar graphs that are directed and those that are unconnected. Since the group of an unconnected graph[3] is given in terms of the groups of its connected components, all we need consider is a procedure for determining the group of a connected graph. In addition, if it is desirable, codes for connected graphs can be generalized to apply to unconnected graphs. Thus codes for trees, to consider one example, can be extended to apply to forests, that is, to an unconnected graph whose components are trees, merely by specifying a unique ordering for the codes of the component trees and separating the tree codes by a distinctive symbol.

With regard to the group of a directed graph, it is clear that an automorphism for a directed graph G is also an automorphism for the associated undirected graph, which is defined as the graph G with the branch directions removed. Thus to determine the group for a directed graph, we apply the algorithm for determining the group to its associated undirected graph. Only automorphisms for the undirected graph are possible automorphisms for the directed graph; however, they may or may not be automorphisms. To complete the test, directions are now reassigned to the branches and the graph with an initial assignment of node numbers is now renumbered in accordance with an automorphic permutation. The incidence matrix for each number assignment is formed and the two matrices are tested for equality; the automorphism holds for the directed graph if and only if the matrices are equal.

This procedure works for planar graphs but not for nonplanar graphs, where directed and undirected graphs represent two distinct problems. The reason for this is that $|\Gamma|$ for the nonplanar graph may be very large; however, for the planar triply connected graph the maximum order is 4b so that the number of matrix comparisons to be made is small and hence manageable. In addition, though $|\Gamma|$ for a series-parallel (S-P) graph may exceed 4b, and may be very large for a tree, directed graphs of these types may still be treated by our algorithm, since incidence matrices need not be formed. Because of the type of unique plane representation of a tree and an S-P graph, determination of whether branch directions can be added without violating automorphism can be done almost by inspection.

As in the preceding paper[1], we consider three basic types of planar graphs, the triply connected graph, the tree, and the S-P graph of connectivity 2 . It is assumed that the reader is aware of the simple algorithm[4] for determining whether a graph is planar, and drawing a plane representation if it is planar. He should also be aware that there are simple tests for determining the connectivity of a graph, and for determining whether it is a tree or an S-P graph[5]. Moreover, it should be mentioned that since dictionaries of graphs based on our codes can now be made, perhaps the easiest way to determine whether a given simple graph is, say, a triply connected graph is to generate its canonic code and look it up in a dictionary of simple triply connected graphs; if the graph is not simple, to determine whether it is triply connected, we first replace each set of parallel branches by a single branch, then determine its canonic code, and finally check whether this code appears in the dictionary.

We now supplement the discussion of trees and S-P graphs given in the previous paper by some comments that are pertinent to automorphism groups. Specifically, we discuss the center and bicenters of a tree in the context of its automorphism group and introduce the concept of a symmetric tree; we also give a detailed example on determining the code and the associated tree of an S-P graph because we make use of this code and tree in determining the automorphism group. It is important to note that the interpretation of the tree associated with an S-P graph is different from that of an ordinary tree.

Since the center C_0 of a tree is unique, it cannot be similar to any other node of the tree; it is thus an invariant node in each automorphism. Correspondingly, since the central branch is unique, its nodes, the bicenters C_1 and C_2, can be similar to no other nodes; they can, however, be similar to each other. This brings us to the consideration of symmetric trees. A tree is said to be symmetric if it has a branch whose two nodes are similar; this branch is called the symmetry branch. It can be shown[6] that a tree possesses either zero or one symmetry branch, and that the symmetry branch, if it exists, is the central branch; thus the nodes of a symmetry branch are the bicenters C_1 and C_2. It is therefore clear that a symmetric tree defines two isomorphic rooted trees, one rooted at bicenter C_1 and the other at C_2. Conversely, two rooted trees that are isomorphic uniquely define a symmetric tree formed from the two trees and an additional branch; one root is made to coincide with one node of the branch and the second root with the other node.

We now give an example to illustrate the determination of the canonic code of an S-P graph and the drawing of a canonic plane representation of the graph and its uniquely associated tree. We consider the series-parallel graph without roots shown in Fig. 1, where the nodes and branches have been labelled. Suppose that we wish to determine the canonic code of this graph without roots; we must test all pairs of adjacent nodes in each of two directions in order to determine the two-terminal graph with the lowest lexical code. As mentioned in the preceding paper[1], we don't actually have to generate codes for many pairs of nodes as terminals. In fact, for this example we need try only one pair of terminals.

First we note that the pairs of adjacent nodes with the highest valences are 1 and 6 and the pair 3 and 4. However, roots 3 and 4 have only two parallel subgraphs whereas roots 1 and 6 have four parallel subgraphs. Since the latter is the maximum of any pair of adjacent nodes, it will yield the lowest lexical code. Pair 1 and 6 are therefore chosen as terminals.

To write the code we must make a choice of the first terminal,for this choice determines the direction in which the subgraphs are considered. It is not necessary to rewrite the code if the choice is the wrong one; as we shall see, simple operations on the code that is obtained will fix the correct choice since the lowest lexical code for either choice and for all possible orderings of parallel subgraphs is determined. For example, by merely taking the mirror-image of the tree corresponding to the generated code, we obtain the code for the S-P graph with its first and second terminals interchanged, more precisely, for the graph rotated 180 degrees. We assume that root 1 is the first terminal; thus our subgraphs are considered from node 1 to node 6.

It is observed by inspection that there are four two-terminal subgraphs between the roots. A component two-terminal subgraph of the most complicated of these subgraphs is the subgraph between nodes 3 and 6. We proceed to consider this subgraph first.

Between nodes 3 and 6 there are two subgraphs in parallel. Starting from branch 3, we have one subgraph consisting of a single branch ℓ in series with the three parallel branches m, n, and o. Its code is therefore $S_2 \ell P_3 mno$; we note that the code would be $S_2 P_3 mno \ell$, if we started from node 6. However, since node 1 is the first terminal, making use of the unidirectionality of S-P graphs, we see that node 3 is the first terminal of this component two-terminal subgraph. Thus we must use the first code. There is no problem with the parallel subgraphs: that is, at this point no order need be specified for them in writing the code and thus any order may be used. Later they will be ordered lexically by means of the tree for the code. The other subgraph from branch 3 to 6 is given by $S_2 P_2 jkp$. The two subgraphs in parallel have a code $P_2 S_2 P_2 jkp S_2 \ell P_3 mno$; however, we could also reverse the order, writing the code with the subgraph containing four branches first: $P_2 S_2 \ell P_3 mno S_2 P_2 jkp$. At this point we do not know, and,in fact, it is not necessary to determine, which will correspond to lowest lexical ordering, since we do not know in which direction we must write the complete code. Now this subgraph is in series with branch i and the total combination is in parallel with h. Proceeding in this way, we determine the complete L-code (that is, L for Lukasiewicz) as

$$L = P_4 S_3 P_2 ab fgcd S_2 e P_2 h S_2 i P_2 S_2 P_2 jkp S_2 \ell P_3 mno \qquad (2)$$

Since we are considering only S-P graphs of connectivity 2, the first character is always an operator P. If we recognize this fact and the fact that the codes are in reduced form, we may show that an equivalent code is given merely by substituting the weights for each of the characters in the code, that is, there is a one-to-one correspondence between the codes. This is true, however, only if the labels of the branches are not significant in the graph; if they do contain information, then we must not destroy the code containing operators and variables and use only the code of the weights.

The code in Eq. (2) shows how to redraw the graph so that the two roots may be put in a horizontal line and the pairs of nodes for other subgraphs in parallel may be drawn so that a left-to-right direction is apparent for them. The redrawn graph is shown in Fig. 2, where parallel subgraphs in the code, in accordance with our convention, have been ordered from top to bottom in the drawing.

A significant correspondence between a rooted tree and the two-terminal S-P graph can now be made; that is, a unique plane rooted tree can represent the code in Eq. (2). Each rooted subtree (except for trivial subtrees) and the tree itself have an operator as the root; its nodes at a unit distance are its associated two-terminal branches or its associated operators. Only free nodes, that is, nodes of valence 1, represent branches of the graph; thus the number of free nodes is equal to the number of branches. Since we can easily determine the two-terminal subgraphs associated with each operator, the corresponding rooted subtrees are made evident.

The tree is drawn in Fig. 3. For purposes of determining the lowest code it is convenient also to have the mirror-image tree, that is, the reflection of this tree in a vertical line; this is shown in Fig. 4. This tree corresponds to the code for the graph in Fig. 2 with interchanged first and second terminals and for the order of parallel branches considered from bottom to top; in other words, it is the code (yielded by the usual conventions of left-to-right and top-to-bottom order) for the graph obtained by rotating the graph of Fig. 2 in the plane of the paper 180 degrees about its center.

Now we can use these trees for determining the canonic code for the two-terminal graph, that is, whether node 1 or 6 should be the first terminal and the order of the parallel two-terminal subgraphs in the canonic plane representation. We should note that, in contrast with the situation[1] where we determined the canonic code for a rooted tree and were free to interchange any pair of branches of a rooted subtree, here we can only interchange nodes at unit distance from a P-operator node; those nodes at unit distance from an S-operator node must remain fixed as originally drawn in the tree and in the mirror-image tree. It is of interest to remark without further discussion because of lack of space that if we also allowed interchange of nodes at unit distance from an S-node, we get a significant result, namely, the canonic code for all graphs 2-isomorphic to the original two-terminal graph; but we are at present interested only in graphs isomorphic to the given graph.

In the trees of both Fig. 3 and Fig. 4, since of all the subtrees of P_4 subtree S_3 has the lowest code and S_2 the next lowest code, we must reorder the subtrees by putting S_3 on the left, then S_2, and finally the two single branches. However, the subtree S_3 for Fig. 3 has the code -2-11111, whereas subtree S_3 in Fig. 4 has the code -211-111. Thus the code for Fig. 3 will have lower lexical rank and node 1 is chosen as the first terminal and node 6 as the second terminal; the tree in Fig. 4 is no longer needed.

We now redraw the tree in Fig. 3 so that it corresponds to lexical order: it is mentioned again that only outward branches from P-nodes may be reordered; the new plane tree is given in Fig. 5. The corresponding code is obtained as usual working up from the subtrees all of whose branches are free branches, is

$$L = P_4 S_3 P_2 abfg S_2 e P_2 S_2 i P_2 S_2 P_2 jkp S_2 \ell P_3 mnohcd \qquad (3)$$

The corresponding S-P graph is shown in Fig. 6, where it is no longer necessary to indicate terminals. This completes the example.

Now a number of theorems may be stated. For example, a pair of two-terminal S-P graphs G_1 and G_2 are isomorphic if and only if their canonic Polish codes are equal. By definition it is required that the first (second) terminal of G_1 correspond to the first (second) terminal of G_2; thus two S-P graphs that are isomorphic may not be isomorphic when these graphs are considered as two-terminal graphs. Also the unique correspondence of two-terminal S-P graphs, the correspondence and S-P graphs to rooted trees leads to a number of theorems and simplifications of treatment of S-P graphs. For example, we may wish to use the X-codes on S-P graphs; to obtain the canonic X-code it is simpler to adopt the convention of using the corresponding canonic plane tree for generating the code. These X-codes can be useful, for example, in considering the group of an S-P graph.

TRIPLY CONNECTED GRAPHS

The algorithm for determining the group of a triply connected graph makes use of the code matrix M. We saw[1] that the rows of this matrix are the 4b X-codes of the graph, where the rows are arranged lexically. The choice of an initial branch and a direction along the branch in the graph or its mirror-image fixes the code that is generated. A branch may be similar to itself or to another branch considered in either of its two directions and with either of its meshes on the right. Thus there are four choices for each branch, and since there are b branches, we get a total of 4b different possible correspondences. Since the maximum possible value of $|\Gamma|$ is 4b, the code matrix M will place in evidence all the automorphisms.

It can be easily shown that the necessary and sufficient condition for two permutations of node numbering to yield an automorphism is that their codes be equal. Thus the matrix M contains one or more sets of equal codes; if there is more than one set, the number of codes in each set is equal. We can now state the following theorem.

Theorem 1. The automorphism group of a planar triply connected graph G is given by the permutations corresponding to the equal codes in any set of equal codes in M. If no codes are equal, the group is the identity.

Thus the algorithm for determining the group requires that we merely determine M and choose a set of equal codes. We then fix the node numbering that corresponds to the lowest code in this set as the original numbering; thus the permutation for this code is the identity. In addition, we now list the permutations corresponding to the other equal codes. These permutations are the elements of the permutation group Γ.

We consider a simple example to illustrate the procedure. We determine the automorphism group of the so-called wheel of order 4, shown in Fig. 7a; its mirror-image graph is given in Fig. 7b. Since $4b = 32$, M will have 32 rows. We find that the code for DA (that is, the code with this branch used as the initial branch in the upward direction) is

$$X(DA) = 12313414525354321 \tag{4}$$

The corresponding node numbering is shown in Fig. 7c. It is clear from the rotational symmetry of the graph that the same code will be obtained for X(A B), X(B C), X(C D) in the original graph and that the corresponding set of four permutations is (1) (2) (3) (4) (5) for the identity, and (1452)(3), (15)(24)(3), (1254)(3) for the other three transformations, where we have used the notation for cyclic permutations. In the mirror-image graph, the same code is given by X(B A), X(A D), X(D C), and X(C B). The permutation corresponding to X(B A) in the mirror-image graph is shown in Fig. 7d which gives (15) (2) (3) (4); the other three permutations are respectively (12)(45)(3), (1)(24)(3)(5), and (14)(25)(3).

Proceeding in a similar manner we find that there are three more sets of equal codes, each containing eight codes. For example, the code for AD in the original graph is X(A D) = 12341451525354321. This is equal to the codes X(DC), X(CB), and X(BA) in the original graph and codes X(DA), X(AB), X(BC), and X(CD) in the mirror-image graph. The node numbering for X(AD) and X(DC) are shown in Figs. 7e and 7f, where we have now omitted the arrows for simplicity. Since previously for Eq. (4) we had Fig. 7c, which gives the correspondence $\begin{pmatrix} A & B & C & D & E \\ 2 & 5 & 4 & 1 & 3 \end{pmatrix}$, and now for Fig. 7e considered as the identity transformation we have $\begin{pmatrix} A & B & C & D & E \\ 1 & 4 & 3 & 2 & 5 \end{pmatrix}$, it is clear that the change in our numbering for the identity is given by $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 5 & 3 & 4 \end{pmatrix}$. The permutation in going from Fig. 7e to Fig. 7f is (4321)(5). This becomes in terms of the number assignment of Fig. 7c (5412)(3), which is observed to be a permutation obtained in the previous group. The reader can check that the same group is obtained for each set of equal codes.

Thus with the nodes numbered as in Fig. 7c, the permutation group of this graph is given by

$$\Gamma = \left\{ \begin{array}{l} (1)(2)(3)(4)(5),\ (1452)(3),\ (15)(24)(3),\ (1254)(3), \\ (15)(2)(3)(4),\ (12)(45)(3),\ (1)(24)(3)(5),\ (14)(25)(3) \end{array} \right\}$$

The example is now complete.

It is worthwhile to mention that it is often possible to eliminate the necessity of generating all the 4b codes when we wish to determine Γ of a given graph. Computation can be decreased significantly by generating only those codes which can possibly be equal. We find these potentially equal codes by checking some necessary conditions that must be satisfied for an automorphism to occur. We illustrate this by an example.

The graph we consider, which is shown in Fig. 8, was also discussed by Ore[2], who mentions that it is difficult to determine that the order of the automorphism group of this graph is two; this fact drops out as a by-product in our check of necessary conditions on the lengths of meshes and the valences of nodes; the results of these checks demonstrate that $|\Gamma|$ is two, and that Γ can be determined by generating only two codes.

Since we desire to use fewer than $4b = 60$ codes in our test, we examine the size of the two meshes that each branch is a part of. A significant fact emerges: branch h is the only branch that is a part of two 5-meshes. Hence we see that h can only correspond to itself and thus the order of the automorphism group cannot exceed four. The further test on necessary conditions considered below decreases the maximum order to two, and then the use of the codes shows that it is precisely two and yields Γ.

At this point we need four codes with h as an initial branch using each of the possible directions in the graph of Fig. 8 and its mirror-image graph. To reduce the required number of codes even further we inspect the length of each other mesh that each branch on the 5-meshes is a part of. In graph $\grave{G}$ the five-meshes, where the faces are encircled clockwise, are given by the ordered sets of branches

$$\begin{array}{c} \{h,\ g,\ d,\ a,\ e\} \\ \{h,\ i,\ n,\ m,\ \ell\} \end{array} \qquad (5)$$

We also observe by inspection that the lengths of the second meshes of the branches (other than h) in the 5-meshes are given by

$$\begin{array}{l} g\colon 4;\ d\colon 3;\ a\colon 6;\ e\colon 4 \\ i\colon 4;\ n\colon 3;\ m\colon 6;\ \ell\colon 4 \end{array} \qquad (6)$$

The mesh branches are considered in order in Eq. (6). However, the order in the mirror-image graph of the mesh lengths would be 4, 6, 3, 4, that is, the reverse of those in Eq. (6); thus the codes for h in the mirror-image graph cannot be equal to those for h in the original graph.

We determine that the code for h = (H, I) in the given graph is

$$X(H, I) = 123451567178929,\ 10,\ 3,\ 10,\ 4,\ 10,\ 98687654321 \qquad (7)$$

We find that the same code is obtained for branch h = (I, H).

Thus obtaining only two equal codes for h, and hence only two equal codes in the code matrix M, proves that the order of the automorphism group of the graph is two. The permutations that are elements of the group correspond to the two equal codes.

Finally, it should be pointed out that the writer has formulated another algorithm[7] for determining the group of a planar triply connected graph; in fact, the algorithm is also applicable to the biconnected graph, and even to the general planar graph. It is not based on the generation of codes but works directly with the plane representations of the graph. Thus a computer technique like SKETCHPAD may be useful in implementing the algorithm; or alternatively, it can be programmed directly without the use of visual displays of graphs.

GROUPS FOR TREES

We may be interested in Γ of a rooted tree or of a tree without a root. Since the treatment of the rooted tree is contained in the discussion of the tree without a root, we need discuss only the latter.

First we must determine a unique rooted tree corresponding to the tree. By the procedure discussed in the preceding paper[1], we determine the center or central branch. If the tree has a center, we use it as the root. We then determine the canonic plane representation and the canonic code of this rooted tree. Since the center is unique, these are of course unique and are defined as the canonic plane tree and the canonic code for the unrooted tree.

If the tree has bicenters, we determine the canonic code for each of the rooted trees, using first C_1 and then C_2 as the root, and then use as the rooted tree the one that gives the lower lexical code. This plane tree and code are defined as the canonic plane tree and canonic code for the unrooted tree. Of course, it is not usually necessary to generate the codes with each of the bicenters as roots; the one to be chosen as the root is often evident by inspection. For example, it should be clear that the one with the higher valence will give the lower lexical code. If their valences are equal, then we inspect the nodes at unit distance from each of the bicenters; the bicenter that has a unit-distance node of higher valence than the valences of the unit-distance nodes of the other bicenter will yield the lower lexical code.

It may occur that each bicenter as root yields the same canonic code; that is, the tree is a symmetric tree. For the definition of the canonic code of the unrooted tree we then use either bicenter as the root. However, for the determination of the automorphism group it is more convenient to delete the symmetry branch and consider the two isomorphic rooted subtrees. Thus, in effect, we are considering a doubly rooted tree, with the bicenters as roots, then deleting the symmetry branch, and interpreting the codes of each rooted subtree as if we had deleted a single root of a rooted tree. An alternative procedure is to consider that the symmetry branch is divided into two branches in series; that is, we have added a node which becomes the root of the rooted tree. Then this case is covered by our treatment of a rooted tree that is discussed below. After the automorphism group of this rooted tree has been determined, we must of course merely remove this invariant node from each of the permutations in the group.

It should be observed, if the codes are used in a dictionary of unrooted trees, that each code, with the exception of a single set of codes, can be inspected to yield its corresponding automorphism group by the simple procedure to be described below; from the above it is clear that the codes for the symmetric trees cannot be interpreted in this way. In other words, symmetric trees must be distinguished in the dictionary; or if they are not, the dictionary user must detect the symmetric trees before generating the automorphism groups. He can do this simply by testing the trees to see whether they contain a central branch and, if they do, whether it is a symmetry branch.

Now we are ready to consider how the group is determined; this may be done by inspection of the canonic code or by inspection of the canonic tree or by consideration of both at the same time. In explaining the procedure it is instructive to work with both the tree and the code. In determining the group of a tree we will make repeated use of the symmetry group $\sigma_n(T_e, T_h, T_j, \ldots)$ operating on the n isomorphic subtrees $T_e, T_h, T_j, \ldots$, where if $n = 1$, we have the identity group. The symmetry group consists of all permutations of the n subtrees and has the order $n!$. We have, in addition, the groups of the subtrees themselves; thus if we now perform these automorphisms after the permutations, we obtain an order for σ_n given by $r^n n!$, where we have used r as the order of $\Gamma(T_e)$, the group of one of the isomorphic subtrees T_e.

The procedure operates on the canonic rooted tree and its rooted subtrees or on the corresponding T-code and corresponding subtree codes. It consists of repeating the basic step of deleting roots and inspecting the resultant set of unconnected trees, that is, the resultant forest. The root of the tree is of course invariant in each automorphism (where, if the tree is a symmetric tree, we are assuming that we use the convention of replacing the symmetry branch by two branches in series and using the added node as the root). The root of a subtree is also invariant in the group of the subtree itself. If the root of the tree that is deleted has a weight of k, where k may be zero or negative, we get a set of 1-k unconnected subtrees. There may be one or more sets of isomorphic subtrees among these 1-k subtrees; such isomorphic subtrees appear in consecutive order in the canonic plane tree and their corresponding subcodes are equal and appear consecutively in the tree code. Some of the isomorphic subtrees may be single nodes.

Thus we can define isomorphism classes for the rooted subtrees. The isomorphic subtrees in each class can be permuted among themselves and after this permutation one can apply an automorphism for each subtree. Thus, as pointed out by Ore[3] in his discussion of a general unconnected graph, the automorphism group of the tree consists of all correspondences written in the form of a generalized permutation which in group-theoretic terminology is called the complete monomial group. However, because we are considering not the general unconnected graph but merely a forest, that is, an unconnected set of trees, it is not necessary to pursue this general approach. The group of the tree, as shown below, is determined quite simply by repeating the basic step discussed above, namely, deleting a root and determining the structure of the group of the resultant forest.

Now continuing the discussion, after the deletion of the root of the tree, we see that it is necessary to find the group of one subtree of each class of isomorphic subtrees, and then permuting the subtrees in each isomorphism class. The required permutation is of course achieved by the symmetry group operating on the subtrees in each class. We then repeat the complete step (of deleting a root and applying the symmetry group to the resultant rooted subtrees in each of the new isomorphism subclasses defined by the root deletion) on a member subtree of the first isomorphism class. The step is applied repeatedly until the automorphism group of the first class has been determined. In carrying out the basic step it is clear that it is applied until an isomorphism subclass of isolated nodes is obtained. We then turn to a member subtree of the second isomorphism class and apply this basic step in the same manner. This is done for all the classes until the group of the original tree is determined.

It is clear that the above can also be carried out by inspection of the code and, if the labels are significant, the corresponding string of labels for the nodes. In the first step we divide the code into its subcodes representing the subtrees. Then we observe the consecutive subcodes that are equal. We then divide one subcode of each set of equal subcodes into its components. We repeat this until we reach subcodes with single characters. The group of each class of equal subcodes is determined by use of the symmetry group of permutations operating on the appropriate subtrees.

We now illustrate this by an example; we will determine the group of the unrooted tree shown in Fig. 9a. The canonic tree is determined and appears in Fig. 9b. Its code is

$$T = -2 - 2 -111 - 1111 - 2 - 111 - 1111 - 111 \qquad (8)$$

The canonic tree is redrawn in Fig. 9c, where the tree and subtree codes are indicated.

Now root i is invariant and thus for each permutation we have $\sigma_1(i)$. We delete root i and obtain three rooted subtrees, two of which have equal codes; thus there are two isomorphism classes and we must continue to operate on subtrees T_e and T_t, shown in Fig. 9d. The group of the tree T_i is thus given by the Cartesian product $\sigma_1(i)\,\sigma_2(T_e, T_\ell)\,\sigma_1(T_t)$ and we have $\Gamma(T_e) = \Gamma(T_\ell)$, by which notation we mean that the two groups are isomorphic.

Deletion of root e in subtree e yields three rooted subtrees, two of which are again isomorphic; in addition, one subtree consists of an isolated node so that no more operations need be performed on it. We thus have for subtree e that $\Gamma(T_e) = \sigma_1(e)\,\sigma_2(T_c, T_f)\,\sigma_1(a)$ and we also have determined that $\Gamma(T_c) = \Gamma(T_f)$. Now considering T_c and deleting root c, we obtain $\Gamma(T_c) = \sigma_1(c)\,\sigma_2(b, d)$. Since b and d are isolated nodes, we have completed the consideration of the first isomorphism class whose subtrees have roots in the 1-level. We then apply the same procedure to the remaining isomorphism class and obtain $\Gamma(T_t) = \sigma_1(t)\,\sigma_2(j, k)$.

Having obtained the structure of the groups of the subtrees in each class, we can now determine the group of the tree. This is done by going back up T_e and determining $\Gamma(T_e)$ and $\Gamma(T_\ell) = \Gamma(T_e)$. First we determine the 2 elements in $\Gamma(T_c) = \sigma_1(c)\,\sigma_2(b, d)$; then by use of $\Gamma(T_c) = \Gamma(T_f)$ we obtain the 4 elements in the Cartesian product $\Gamma(T_c)\,\Gamma(T_f)$; next we obtain the 8 elements in $\sigma_2(T_c, T_f)$, which finally yields the 8 elements in $\Gamma(T_e) = \sigma_1(e)\,\sigma_2(T_c, T_\ell)\,\sigma_1(a)$. We also now can write the 64 elements in the Cartesian product $\Gamma(T_e)\,\Gamma(T_\ell)$ and thus the 128 permutations in $\sigma_2(T_e, T_\ell)$. Finally, we obtain $\Gamma(T_i) = \sigma_1(i)\,\sigma_2(T_e, T_\ell)\,\sigma_1(T_t)$, which is seen to possess 256 elements.

As mentioned previously, we can also operate on the code directly. We delete the first operator and obtain the division into subcodes operated on by this first operator as $\underline{\text{-2-111-1111}}$ $\underline{\text{-2-111-1111}}$ $\underline{\text{-111}}$. We note the two subcodes equal to -2-111-1111. We thus must consider this subcode and the last one, -111. We now operate on the subcodes, first considering -2-111-1111. We delete the first operator and obtain the subdivision $\underline{\text{-111}}$ $\underline{\text{-111}}$ $\underline{\text{1}}$. The process should now be clear. At each step we record the part of the group that has been determined up to that point. In this way the group of the original tree is found.

GROUPS FOR SERIES-PARALLEL GRAPHS

As in the case for the rooted and the unrooted tree, separate discussion of the two-terminal S-P graph and the S-P graph is not required; the former is included in the treatment of the latter.

It would appear at first thought, since a unique rooted tree can be associated with an S-P graph, that the previous discussion of the rooted tree also covers the S-P graph. However, this is not quite true; careful interpretation of the associated tree is required. Since tree nodes that are not free nodes represent operators, we must determine what isomorphic rooted subtrees whose roots are operators tell us about the automorphic node permutations in the original S-P graph. In addition, the free nodes of the associated tree represent not nodes but branches of the S-P graph; thus interpretation is also required here.

Since the S-P graph may yield the same canonic code when it is rotated 180 degrees and its parallel subgraphs are reordered to give lowest lexical rank, we must consider not only the associated tree but we must know whether the first and second terminals may be interchanged; this information is of course generated when we find the canonic code. Finally, if any other pair of terminals yields a code equal to the canonic code, we will also have this information produced by the determination of the canonic code. On the other hand, if we are given the code or the associated tree of the S-P graph and not the graph itself, and the information on the equal codes is not given, we must first check whether equal codes exist. This may be done very simply; for example, as is clear from the discussion of Fig. 1 in the Introduction, to determine whether interchanging terminals will yield an equal code, we consider the mirror-image graph of the associated tree and reorder each of its subtrees with a P-operator as root so that the lowest lexical code is obtained.

Furthermore, in searching for equal codes we must only determine equal canonic codes; the equal codes that are not canonic merely give redundant information obtained from the equal consecutive subcodes (of length > 1) of the canonic code. For example, for the S-P graph shown in Fig. 10a the canonic code is obtained using nodes a and b as terminals 1 and 2, respectively; equal canonic codes are obtained using terminals a and c, terminals a and e, and terminals a and d. All these equal codes must be considered in determining the group. It should be noted that though nodes a and f will yield a lower lexical code, they are not proper terminals since they are not adjacent. For the graph in Fig. 10b, however, the canonic code is obtained using terminals a and f; the equal noncanonic codes obtained using as terminals a and b, a and c, a and d, and a and e need not be considered, since the automorphisms placed in evidence by the equal codes are already taken into account in the equal consecutive subcodes of the canonic code. We can see that this will be true in general for the following reason: when the canonic code is obtained using as terminals one or two internal nodes of a subgraph that is identical to one or more other subgraphs in parallel with it and thus can be permuted with these subgraphs to yield automorphisms, then equal canonic codes will be obtained using other corresponding pairs of terminals in these subgraphs; however, when the canonic code does not use as terminals any internal nodes of the parallel set of identical subgraphs, this parallel set will be in series or parallel with other subgraphs and hence will have the automorphisms obtained by permuting its identical subgraphs made evident by equal subcodes of length > 1 in the canonic code.

Now we are in a position to interpret the canonic code and the associated tree in order to determine the automorphism group. We follow a procedure similar to that for determining the group of a tree; however, since only parallel subgraphs may be permuted, we permute only the subgraphs corresponding to the isomorphic subtrees obtained when a root that is a P_i-operator is deleted; in the code this means we look for equal consecutive subcodes that are among the i subcodes operated on by P_i. When an S_i-operator is deleted and yields isomorphic subtrees, we form isomorphism classes as before, but we do not permute the subgraphs corresponding to the subtrees in an isomorphism class. When parallel subgraphs are permuted, the symmetry group operates on corresponding internal terminals of the parallel subgraphs. For example, when deletion of a root P_3 yields two isomorphic rooted trees with S_3 as root and one isolated node, as eill be true for the graph in Fig. 11, then internal terminals for the three two-terminal subgraphs in series, namely terminals a and b, and terminals c and d, each pair considered as an ordered pair, are permuted. Whether the external terminals of the parallel subgraphs may be permuted is determined by equal canonic codes; for example, in this S-P graph either e or f may be the first terminal for the canonic code, that is, we obtain equal codes, so that nodes e and f are similar.

Using the above discussion and the discussion in the preceding section on the group of a tree, the reader should now be able to complete the example by obtaining the code, the associated tree, and finally the group of the S-P graph in Fig. 11.

CONCLUSION

Based on codes previously formulated by the author[1] algorithms were given for determining the automorphism groups of planar graphs, specifically, the groups of triply connected graphs, trees, and S-P graphs of connectivity 2. The discussion is given for undirected graphs but it is pointed out how the procedures may be applied to directed graphs. A combination of the procedures may also be applied to determine the group of a general planar graph; this will be discussed in a future paper. We also hope to treat nonplanar graphs by use of similar concepts.

REFERENCES

1. Weinberg, L., "Plane Representations and Codes for Planar Graphs," presented at the Third Allerton Conference on Circuit and System Theory, Oct., 1965, published in these Proceedings.
2. Unger, S.H., "GIT-A Heuristic Program for Testing Pairs of Directed Line Graphs for Isomorphism," Comm. ACM, Vol. 7, No. 7, No. 1, Jan. 1964, pp. 26-34.
3. Ore, O., Theory of Graphs, Amer. Math. Soc. Colloq. Public, Vol. XXXVIII, Prov., R.I., 1962.
4. Goldstein, A.J., "An Efficient and Constructive Algorithm for Testing Whether a Graph Can Be Embedded in the Plane," unpublished memorandum, BTL, Murray Hill, New Jersey.
5. Weinberg, L., "Linear Graphs: Codes, New Theorems, and Algorithms for Determining their Isomorphisms and Automorphism Groups, with Applications," to be published as a Univ. of Mich. Radar Laboratory Report.
6. Otter, R., "The Number of Trees," Ann. of Math., Vol. 49, No. 3, July 1948, pp. 583-599.
7. Weinberg, L., "An Alternative Algorithm for Determining the Automorphism Group of a Planar Graph," to be submitted for publication.

Fig. 1 Series-parallel graph with labelled nodes and branches.

Fig. 2 Redrawn graph with parallel branches ordered in accordance with the code.

Fig. 3 Tree corresponding to the code in Eq. (2) for the graph in Fig. 2.

Fig. 4 Mirror image of tree in Fig. 3, corresponding to code for graph of Fig. 2 rotated 180° so that node 6 is the first terminal and node 1 is the second terminal.

Fig. 5 Lexically ordered tree isomorphic to the tree in Fig. 3.

Fig. 6 Plane two-terminal graph with lowest lexical code.

Fig. 7 Graph for which the automorphism group is found, and some related graphs.

Fig. 8 Graph for which the automorphism group is to be found.

Fig. 9 Unrooted tree, its canonic plane representation, and two of its subtrees, where codes of the tree and all subtrees are indicated.

Fig. 10 Two S-P graphs, each of which has some equal codes among the codes for its two-terminal S-P graphs.

Fig. 11 S-P graph with two identical subgraphs in parallel.

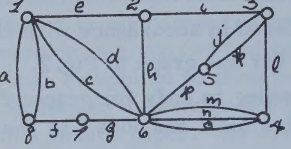

1
2
3
4
5
6
7
8
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p

FIGURE 1

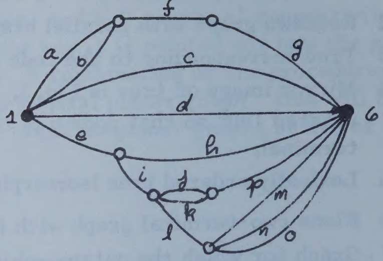

1
6
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p

FIGURE 2

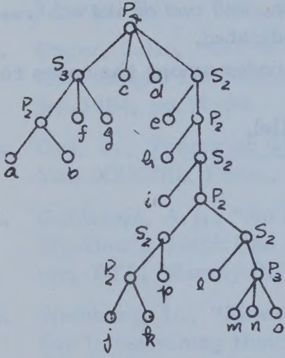

P4
S3
S2
P2
c
d
e
f
g
a
b
h
i
j
k
l
m
n
o
p
P3

FIGURE 3

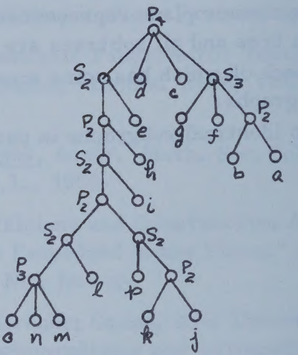

P4
S2
S3
P2
P3
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p

FIGURE 4

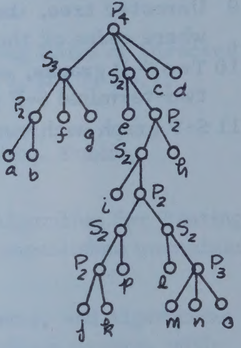

P4
S3
S2
P2
P3
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p

FIGURE 5

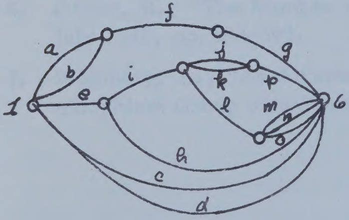
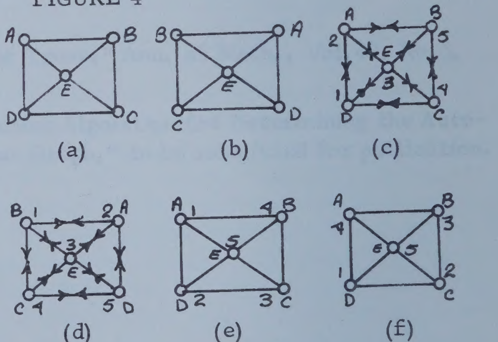

1
6
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p

FIGURE 6


A
B
C
D
E
1
2
3
4
5

(a) (b) (c)

(d) (e) (f)

FIGURE 7

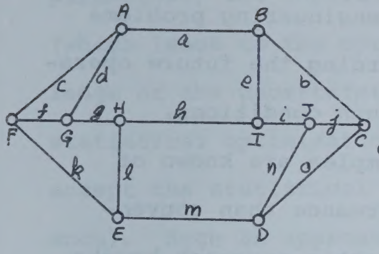
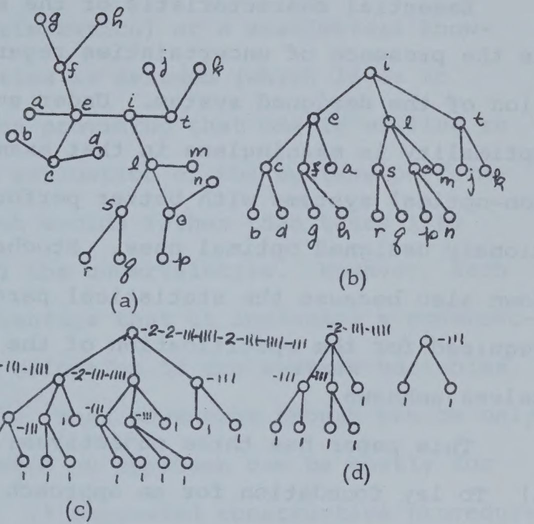

A
B
a
c
d
e
b
F
f
G
g
H
h
I
i
J
j
C
k
l
n
o
m
E
D

FIGURE 8


g
h
f
j
k
a
e
i
t
b
d
m
c
l
n
s
o
r
q
p
i
e
l
t
c
f
a
s
o
m
j
k
b
d
g
h
r
q
p
n
-2-2-111-1111-2-111-1111-111
-2-111-1111
-2-111-1111
-111
-111
-111
-111
-111
-2-111-1111
-111
-111
-111

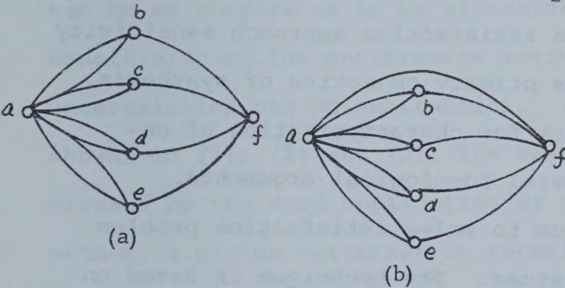
(a)

(b)

(c)

(d)

FIGURE 9


b
c
a
d
e
f
b
a
c
d
e
f

(a)

(b)

FIGURE 10

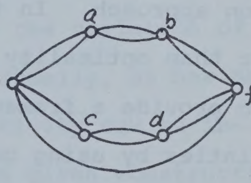


FIGURE 11

SATISFACTION APPROACH TO THE SYNTHESIS AND CONTROL OF SYSTEMS

MIHAJLO D. MESAROVIC
Systems Research Center - Case Institute of Technology
Cleveland, Ohio

Essential characteristic of the engineering problems is the presence of uncertainties regarding the future operation of the designed system. Under such conditions, optimality is meaningless in that examples are known of non-optimal systems with better performance than conventionaly designed optimal ones. Stochastic approach breaks down also because the statistical parameters which are required for the specification of the system are themselves unknown.

This paper has three objectives:

1) To lay foundation for an approach to deal with the problem of synthesis under uncertainty, so called satisfaction approach. In the satisfaction approach sensitivity rather than optimality is primary objective of synthesis.

2) To provide a framework for characterization of uncertainties by using general topological arguments.

3) To present a technique to solve satisfaction problem for a class of linear systems. The technique is based on the convex programming approach and can also be represented as being an approximation to the game-theoretic solution of the problem.

1. INTRODUCTION

There exists a definite need for the broadening of the conceptual basis for the formal, mathematical, theories of synthesis and control of systems. This, in particular, refers to the treatment of the uncertainties which are always present in the engineering problems. Traditionally, in the process of selecting the systems variables or parameters, the uncertainties are either completely ignored (which leads to the optimization) or a statistical knowledge of the uncertainties is assumed (which leads to statistical optimization providing that one is willing to accept the statistical evaluation of the systems performance). Such an approach avoids rather than takes into account (or eliminates) the uncertainties. However, such an approach has the advantage that it indicates a constructive procedure for the selection of the systems variables (e.g. via optimization). This advantage though can be only apparent and applying such an approach can be costly for the following reasons: 1) Suggested constructive procedure can be so complex as to be effectively non-realizable; 2) Sensitivity of the performance versus the variation of the uncertainties can be substantial. (Actually, as has been shown, in (1), (2) and (3), the sensitivity can be increased by the very application of the given constructive method, i.e. the optimization technique.)

In this paper we shall be concerned with the second problem.

Traditionally, the problem of dealing with the uncertainties is approached via adaptation which results in a two-step procedure:

1) The selection of variables or parameters is made by ignoring the uncertainties and the selection is implemented on actual system; 2) Behavior of the system (with the

implemented selection) is observed and the assumptions about the systems operating conditions are modified so that at the next decision time the selection is improved.

In this paper we shall propose, and to some extent explore, the possibilities of an alternative approach of dealing with uncertainties. The presence of the uncertainties will be accepted and taken into account in the formulation of the selection process itself. An on-line adaptation still might follow the selection but its objective now is changed in the sense that the prediction of a class of uncertainties (e.g. via its characteristic pattern or class property) is involved rather than the prediction of a point (parameter or variable) in a space. The approach is based on the earlier papers

2. FORMAL DEFINITION OF THE PROBLEM

To focus the attention on the essential elements of the problem it will be formulated in the set-theoretic context, in the spirit of the general theory of systems (4) and (5).

To define the problem of selecting a system variable the following has to be given:

1) Set of alternatives $M = \{m\}$

2) Systems function $S: \; M \times U \rightarrow Y$

3) Uncertainty set $U = \{u\}$

4) Performance functional $G: Y \times M \rightarrow Q$ where $Q = \{q\}$

5) Tolerance functional $T: U \rightarrow A$ where $A = \{\alpha\}$. (In more general cases the mapping T is $U \times M \rightarrow A$)

6) Satisfaction relation $R \subset Q \times A$

The problem is then:

Find $m \in M$ such that $[F(m,u), T(m,u)] \in R$ for all $u \in U$

where

$F: U \times M \rightarrow Q$ is derived from G and S, i.e.

$S(M \times U) \times M \rightarrow Q$.

If A is a unit set, and R is an ordering relation i.e. $[(q,\alpha) \in R] \rightarrow (q \leqslant \alpha)$, element $\alpha \in A$ can be given a simple interpretation of a tolerance limit and the problem is then to find $m \in M$ such that the tolerance limit is not violated for any of the uncertainties, $u \in U$. This is responsible for calling the problem a satisfaction problem since the goal is to find $m \in M$ which will satisfy the relation R.

It is interesting to note that a school of thought in the psychology of human decision-making argues that the human decision-making behavior can be best explained by assuming that he strives to satisfy certain minimal requirements, that essentially he is looking for a good-enough solutions and will not exert himself to further pursue a solution up to a performance which is an optimal one (9). Such a behavior can be described by the formal satisfaction problem presented here.

Notice that the optimization problem is obtained as a special case of the satisfaction problem by assuming: a) Uncertainty set is a unit set; b) Tolerance limit α is given implicitly by requiring $\alpha = \min q$; c) Satisfaction relation R is the equality: $q - \alpha = 0$.

The principle weakness of the satisfaction problem is that in general, it does not have a unique solution and the formulation does not indicate how one can proceed to find one of the alternative solutions. However, it should be realized that this difficulty is not due to the formulation of the problem but is a very essential characteristic of the real-life problem of which our problem is merely a formalization. It is, therefore, only appropriate that this basic aspect is preserved under formalization.

3. TOPOLOGICAL SPECIFICATION OF THE SATISFACTION PROBLEM

To develop a more constructive theory of the satisfaction problem, one should introduce more structure in the elements specifying the problem so that the properties of the solution can be explored yet the problem is not changed beyond recognition from the original real-life problem.

Here, we shall have time and space only to indicate how the problem is changed by the introduction of the general topological structures without going into the study of the properties of the solution due to the newly introduced structure.

The most natural way to introduce topologies is via extended topology of P. Hammer (10). Consider the uncertainty set U. Let v be a function mapping power set P(U) into itself, $v: P(U) \rightarrow P(U)$, and let v be an expanded function i.e. for any two subsets $X, Y, \epsilon P(U)$, $X \subseteq Y$ implies $v(V) \subseteq v(Y)$ and furthermore $v \supseteq e$ where e is the identity function in P(u). Let c be the complement function in P(U) i.e. for any $X \in P(U)$, $c(X) = P(U) \sim X$. We can now define a neighborhood function in U by means of v and c.

A set $X \subseteq U$ is a v-neighborhood of a point $p \in U$ iff $p \in c\ v\ c\ (X)$. Apparently every point and every subset of U has a neighborhood. Pair $\{U, v\}$ is now an extended topological space. Notice that if v has additional properties of closure, it becomes Kuratowski closure function and the pair $\{U, v\}$ becomes a usual general topological space. These additional properties are not necessary here and as it has been shown in (7) they can be rather unnatural for some approximation problems.

Using the notions of a v-neighborhood we can now redefine the satisfaction problem so as to allow for a less rigid, approximate, solution.

Consider the set A and let (A, T_A) be an extended topological space. Take a subset C T_A such that C is a cover for A. We can formulate now the satisfaction problem in the following way:

Given a pair $m \in M$ and $U' \subset U$. m will be considered a satisfaction solution in U' with respect to θ iff for any $u \in U'$ there exist a pair $\alpha \in A$ and $\theta_\alpha \in 0$ such that $[F(m,u), \alpha] \in R$, and θ_α is a neighborhood for both α and $T(m,u)$ i.e. $(\alpha \in \theta_\alpha) \wedge [T(m,u) \in \theta_\alpha]$

In an important special case T is a constant mapping (i.e. only one α is acceptable) and furthermore Θ is a partition; $m \in M$ is then a satisfaction solution in U' iff for every $u \in U'$ there exists a $\alpha_u \in \theta_\alpha$ such that $[F(m,u), \alpha_u] \in R$; where θ_α is the v-neighborhood of α. Subset is then a measure of tolerance for the solution of the satisfaction problem in the sense that if the performance q satisfies R with any element $\alpha_r \in \theta_\alpha$ the solution is considered as satisfactory.

Topology in A induces in a natural way a topology in U. Consider the case when $Q = A = N$ and N is linearly ordered by $<$. For any pair (m,α) there is defined a subsets of U

$$U_{\alpha,m} = \{u : F(u,m) < \alpha\}$$

It can easily be shown that the family $L = \{U_{\alpha,m}\}$ is a general topology for U. Furthermore, let T be a constant mapping $T : U \times M \rightarrow \{\alpha\}$ where $\{\alpha\}$ is a unit set and the satisfaction relation R be given by

$$R = \{(q,a) \ : \ (a = \alpha) \wedge (q < \alpha)\}$$

To every $m \in M$ corresponds a subset $U_{\alpha,m}$. A solution m_i will be called preferable to another m_j if

$$U_{\alpha,m_j} \subset U_{\alpha,m_i}$$

This condition has two important implications:

1) It indicates that a preferable solution has to satisfy the requirements in a larger subset of U. In other words, a preferable solution will not violate the constraints in a broader class of uncertainties. This is equivalent to stating that a preferable solution is less sensitive in U.

2) By taking m_j to represent the behavior of the system S with an arbitrary selection of m. The condition shows that a satisfactory element $m_i \in M$ has to improve the sensitivity of S with respect to the variations in U. The objective in the solution of the satisfaction problem, therefore, is the improvement in the sensitivity of the system.

4. CHARACTERIZATION OF THE UNCERTAINTY SET

A major problem in the formulation of the satisfaction problem is the characterization of the uncertainty set U. The difficulty in defining U in any given case stems from the fact that the set U reflects the entire ignorance one has about both the environment and the system itself. Actually, U embodies all the hypothesis about the system and environment which the designer or the decision maker makes in the formulation of the problem.

To get some analytical results, a more special, letter structured, representation of U has to be given. A particular convenient way is to represent U as a linear vector space. By using the well known inequalities from the functional analysis, one can study then the existence of the satisfaction element $m \in M$ and the bounds on U for which such an element exists.

As an illustration, let U be a normed linear space and the satisfaction problem be defined by: a) S as a linear operator; b) G as the inner-product operator such that

$$Q = \{q:q = q(m,u) \wedge (q(m,u) = [(y-r),A(y-r)] + (m,B_m)\}$$

Consider the problem of finding a bound in the normed space U which will guarantee the solution of the satisfaction problem. Let m be the selected element; $\|u\|$ the norm of the uncertainty $u \in U$; y_o the element of Y describing the behavior of the system if no control is applied (so called free motion of S); r the variable from the inner product form (reference variable).

Given an element of the uncertainty set $u' \epsilon U$(e.g. u' can be obtained by prediction). Let $m_p \epsilon M$ be such that for all $m \epsilon M$

$$q(m_p, u') \leqslant q(m, u')$$

i.e. m_p is an optimum. Define the tolerance limit as a functional on u such that for any $u \epsilon U$, $\alpha(u) = q(o,u)$, i.e. α is the value of the performance when the control is not applied. Finally, define the satisfaction relation by inequality such that for any

$\alpha(u) \epsilon A$, $[q, \alpha(u)] \epsilon R$ implies $q \leqslant \alpha$.

For this problem it has been shown (8) that the satisfaction relation is satisfied in the subset U'

$$U' = \{u: \|u - u'\| \leqslant 1/2 \| y_o + u' - r \| \}$$

The result indicates that there exists a relation between the deviation of the output variable from the reference variable on one hand and the maximum norm of the uncertainty set U' in which the systems performance is sufficiently insensitive, on the other. It indicates the advantage of applying the control variable in terms of sensitivity.

5. EQULIZATION HEURISTIC FOR LINEAR CONTROL PROBLEMS

Satisfaction problem does not have, in general, a unique solution, i.e. there exists a subset $M' \subset M$ whose elements satisfy the relation R.

Conceptually, the simplest ways to change the satisfaction problem so as to provide for a unique solution are the following:

a) To require that the selected m, in addition to satisfying R, also optimizes a functional (perhaps q itself) in a given sense. The problem becomes then an optimization with the satisfaction problem defining the hard constraints.

b) To require the strictest tolerance, i.e. require the minimal value for the maximum α over the entire U. The satisfaction problem becomes then a game-theoretic problem, where the satisfaction relation is usually satisfied even in a larger subset of U than originally required.

It is important to realize that the conditions introduced in a) or b) are additional conditions not contained in the specification of the original problem but introduced subsequently to make the problem conceptually simpler. If that leads to a simple solution, the procedure is certainly justified. Unfortunately, this is rarely the case especially for the second, i.e. game-theoretic approach. Additional conditions make the problem practically unsolvable and the conceptual nicety of having a unique solution is of no avail. An alternative approach then has to be explored. Instead of looking for a general solution, a special class of problems has to be considered and on the basis of the familiarity with the structure of the problem, a procedure can be suggested which will solve that particular type of problems but will have little power in more general cases. Such a procedure, if applied outside of this narrow class, is called a heuristic. To increase the confidence in the procedure, the conditions should be explored in which the heuristic precisely solves the problem exactly, i.e. under which the procedure is a solution algorithm. Such an approach to the solution of the satisfaction problem might not have a estetic attractions of the approaches a) or b) yet its principal virtue lies in the possibility in arriving at a simple solution algorithm on which none of the more elaborate procedures (a or b) can improve.

To illustrate the possibilities for solving the satisfaction problem by such heuristic approach considered the following problem:

Let M,Y and Y be linear vector spaces and the system S be defined by the difference equations.

$$\overline{y}_{j+1} = A\,\overline{y}_j + B\,\overline{m}_j + C(\overline{u}_j)$$

Let the performance functional G be given by the inner product

$$Q = \{q : q = \Sigma[\overline{y}_j\overline{y}_j + \overline{m}_j\,D\,\overline{m}_j]\}$$

A,B,C and D are constant matrices.

Finally, let the satisfaction relation be given by the inequality

$$R = \{(q,\alpha) : q \leq \alpha\}$$

where α is a constant.

Satisfaction problem consists then in finding the vector sequence $\overline{m} = \{\overline{m}_1, .. \overline{m}_N\}$ such that R is satisfied for all $\overline{u} \in U$ where $\overline{u} = \{\overline{u}_1, ... \overline{u}_N\}$.

Let $\overline{u}$, $\overline{m}$ and $\overline{y}$ be K-dimensional vectors. Set of solutions of the satisfaction problem $M' \subset M$ form then a closure of the intersection of 2^{NK} ellipsoids in the space U. Study of M' (3) shows that its maximum diameter is inversely dependent on the diameter of the uncertainty set.

Let $U' \subset U$ be the region in which the satisfaction relation R holds. The region U' much like M' represents an intersection of the second-order hyper-surfaces.

Using these special properties of the problem the following procedure is then suggested as a heuristic to solve the satisfaction problem:

Find $\overline{m} \in M$ such that the performance functional q has the same value as the opposite extreme corner points of U'.

This procedure is called equilization heuristic.

Detailed study of the equilization heuristic (3) shows that it has the following properties:

1) It leads to a simple decision rule ("control law") which is a linear function of the states.

2) For a broad class of problems if the equilization heuristic does not have a solution then none exists.

3) If some conditions on the systems properties are satisfied the equilization heuristic results in the same $\overline{m}$ as the game theoretic approach. In other words, the heuristic leads to the minimal tolerance. (This also indicates that the heuristic can be used for solving some types of games).

4) The heuristic results in a system which is less sensitive in large than the system obtained by the minimization of the expectation of the performance functional. Furthermore, the equilization heuristic minimizes a measure of sensitivity in the small.

6. CONCLUSION

Solution of the synthesis and control problems should start with a formalization which is as close to the real-life situation as the formal apparatus allows. This leads to the satisfaction problem proposed in this paper. To make the progress toward the solution of the satisfaction problem, it is necessary to possess a proper characterization of the uncertainty set U. One of the most important problems in the area of synthesis and control is then how to characterize the uncertainty set. Hopefully, this can be done so that some analytical results can be obtained. However, judging from the complexity of the satisfaction problem, it seems that the analytical method might be used to investigate properties of various solutions while the actual search for a specific solution has to be done with the computer which should be involved not only in providing the numerical answers but as a part of the logical process in the problem solving.

REFERENCES

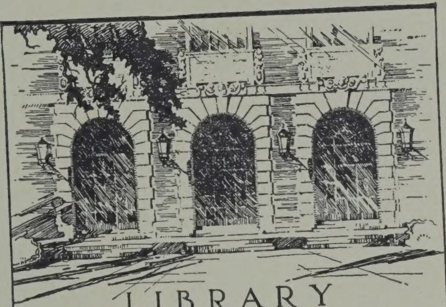
1. Macko, D.S. and Mesarovic, M.D., "Uncertainties and Optimal Control Approach to Feedback Control Problems", Information & Control, (1965)
2. Culver, W., "A Modeling Theory for a Class of Optimal Control Systems", SRC 38-A-63-15, (1963)
3. Markarian, M.D., "A Heuristic For Decision Making Under Uncertainty", SRC 74-A-65-27, (1965)
4. Mesarovic, M.D. (Editor), "Views on General Systems Theory", John Wiley, (1964)
5. Windeknecht, T.G., Mesarovic, M.D., "On General Dynamical Systems and Finite Stability", Proceedings of the Conference on Dynamical Systems, Puerto Rico, (1965) To Appear
6. Simon, H., "Models of Man", John Wiley, (1966)
7. Hammer, P., "Extended Topology", University of Wisconsin Reports
8. Takahara, Y., "Multi-Level Systems and Uncertainties", To appear

INDEX OF AUTHORS

Aggarwal, J. K. 514
Ahmed, N. 817
Anderson, D. R. 524
Auth, L. V. ,Jr. 906
Bacon, C. M. 841
Baker, T.S. 130
Bakker, M. L. 840
Bedrosian, S. D. 745
Bickart, T. A. 283
Bingham, J.A.C. 773
Bingulac, S. 212,714
Biswas, R. N. 384
Blostein, M. L. 421
Brown, D. P. 827
Burzio, A. 230
Butler, E. M. 845
Calaham, D. A. 394
Carlin, H. J. 14
Chan, S. P. 99
Chen, W. H. 546
Chen, Wai-Kai 889
Christian, E. 806
Chua, L. O. 502
Cruz, J. B., Jr. 607
Darlington, S. 7
Deo, N. 879
Desoer, C. A. 259
Dicker, D. 144
Dobbs, D. H. 844
Dorato, P. 35
Dorf, R.C. 46,225
Edenhofer, J.R. 726
Eisenmann, E. 806
Golden, R. M. 690
Goodnuff, J. L. 772
Graefe, P.W.U. 270
Gulati, J. 745
Haase, K. H. 846
Hakimi, S. L. 879
Halijak, C. A. 856
Hall, W. B. 772
Ham, J. M. 578
Hassun, R. 867
Hausmann, E. C. 844
Hemesath, N. B. 362
Ho, B. L. 449
Hollister, F. H. 430
Holzmann, C. A. 536

Huang, T. S. 722
Huber, E.A. 182
Ishizaki, Y. 786
Jenson, H. 868
Johnson, D. E. 827
Johnson, J. R. 827
Kailath, T. 320
Kaiser, J. F. 621
Kalman, R. E. 449,563
Karni, S. 817
Katzenelson, J. 647
Kestenbaum, A. 35
Kim, H. K. 172
Kokotovic, P. 27,212
Ku, W. H. 722
Kuh, E.S. 384
Kuo, F. F. 192
Kurz, L. 329
Lago, G. V. 844
LaPatra, J. W. 868
Larson, R. E. 472
Lasdon, L. S. 677
Lefschetz, S. 1
Lee, C. T. 259
Lee, H.B. 715
Lee, S. C. 407
Leon, B. J. 502
Levy, D. M. 827
Lin, P. M. 826
Liou, M. L. 669
Liu, R. 73,536
Loh, N. K. 270
McBride, L. E. 492
McKinney, J.A. 856
Marshall, T. G., Jr. 200
Massey, J. L. 73
Masten, L. B. 857
Maxwell, L. M. 727
Medanic, J. 212
Meier, L., III 109
Meissinger, H. F. 589
Mesarovic, M. D. 930
Meyer, E.R. 99
Mitra, S. K. 845
Morgan, B.S., Jr. 252
Mowle, F. 82
Mueller, T. E. 339
Murata, T. 751
Murray-Lasso, M. A. 201
Myers, B. R. 906
Nardizzi L. 613
Narendra, K. S. 130
Olivares, J. E., Jr. 513
O'Neil, P. V. 900

O'Neill, W. D. 353
Onaga, K. 762
Papoulis, A. 202
Perkins, W. R. 607
Peschon, J. 472,482
Porter, W. A. 241
Pottle, C. 659
Protonotarios, E. N. 162
Puri, N. N. 369
Reeder, J. R. 310
Rekasius, Z.V. 439
Richie, C. G. 514
Rutman, R. S. 27
Sablatash, M. 796
Saeks, R. 727
Sage, A. P. 828
Sain, M. K. 121
Schaefgen, H. W. 492
Scheibe, P. 182
Seacat, R. H. 857
Seamon, J. H. 843
Seitelman, L. H. 647
Sharpe, G. E. 512
Shenoi, B. A. 151
Siljak, D. D. 225,230
Simon, M. K. 329
Slepian, P. 900
So. H. C. 634
Steele, S.A. 282
Steiglitz, K. 492
Stephenson, S. E. 842
Sullivan, D. 82
Tahim, G.S. 346
Tang, D. T. 63
Temes, G. C. 773
Thaler, G. J. 430
Thathachar, M.A.L. 130
Thomas, J. B. 299
Tomovic, R. 17
Trick, T. N. 524
Triska, C. J. 840
Wagner, T. J. 91
Wakazuki, H. 786
Waren, A. D. 677
Watanabe, H. 786
Weinberg, L. 733,913
Wing, O. 162
Wohlers, M. K. 708
Wolff, S.S. 310
Womack, B. F. 843
Yao, K. 299
Zames, G. 460



LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
3 1965
3 0112 006279605